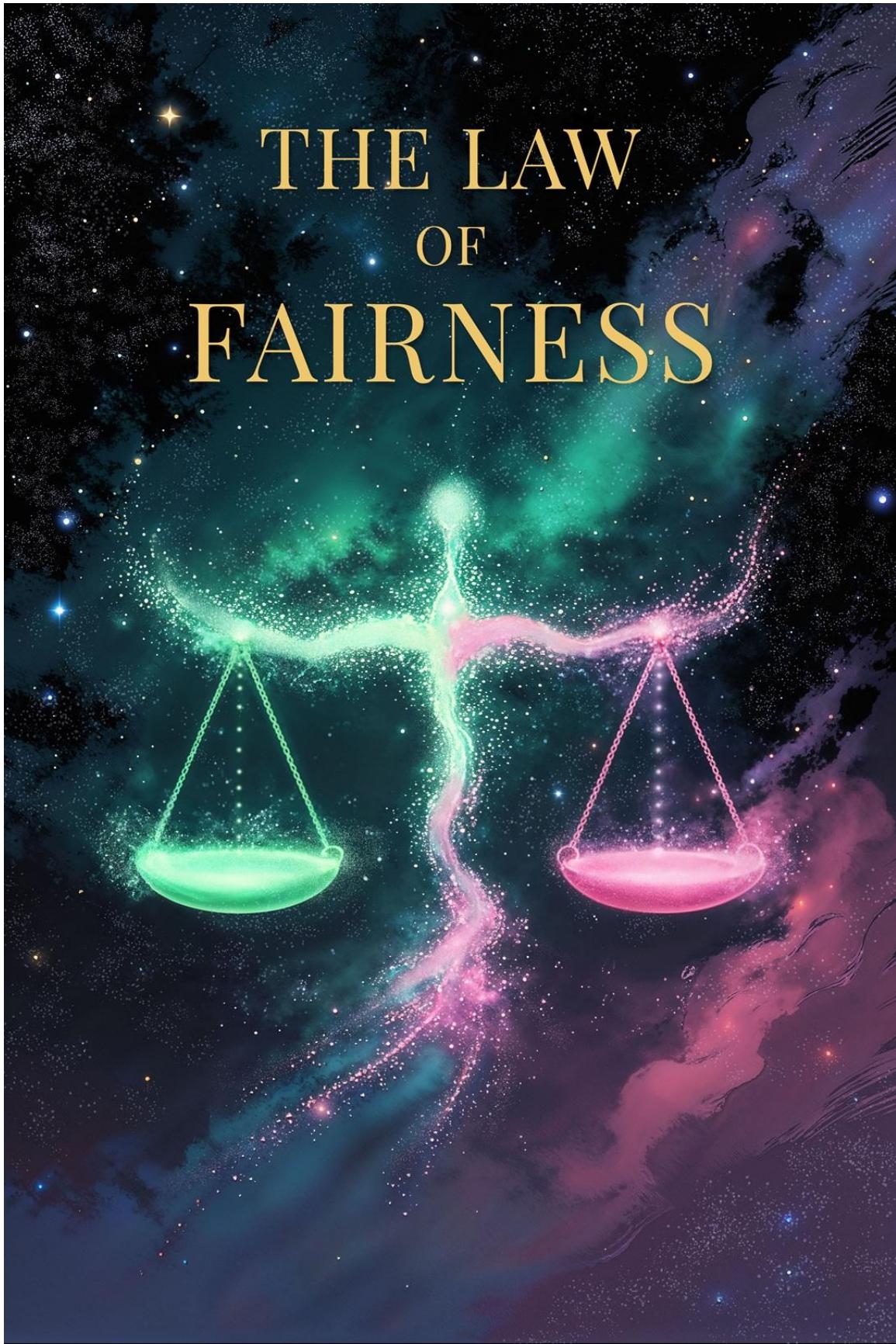


THE LAW OF FAIRNESS



THE LAW OF FAIRNESS

Table of Contents

Table of Contents	3
Illustrated Guide to Fairness	10
Part I — The Question That Won’t Go Away.....	32
Chapter 1 — Why Fairness? Why Now?	35
1.1 A Child, a Rockstar, a Monk	39
1.2 What People Usually Mean by Fairness	41
1.3 The Hard Cases No One Likes.....	45
1.4 Why “Tendency Toward Fairness” Isn’t Enough.....	49
1.5 How This Book Works.....	52
Chapter 2 — Feelings as the Final Currency	54
2.1 Why Feelings, Not Just Things	57
2.2 Can Feelings Be Counted?	60
2.3 Somatic Markers	66
2.4 Affective Neuroscience in One Page.....	72
2.5 What We Will Never Do	78
2.6 Preview: Hedonic Composite Units (HCU) and the Life Ledger	84
Part II — The Law, Stated Clearly	89
Chapter 3 — The Law of Fairness	92
3.1 Canonical Statement.....	95
3.2 Six Assertions of the Law.....	100
3.3 What the Law Does Not Say.....	105
3.4 Boundary Conditions	108
3.5 The Death of Mind.....	120
3.6 Research Notes: The Ledger Integral and State–Change Formalism	128
Chapter 4 — Constraint, Not Purpose.....	143
4.1 Guardrails vs. Steering	146
4.2 Why Constraints Beat Miracles	151
4.3 No Teleology	155

4.4 Lawhood: Best System vs. Governing Law	160
4.5 Research Notes: Optional Stopping and Regularity	168
Part III — How the System Works (From the Inside)	177
Chapter 5 — The Queue System (QS)	182
5.1 Queue System in a Sentence	187
5.2 Choice Sets and Admissible Policies.....	192
5.3 Neural Correlates: rIFG, ACC, vmPFC, Insula.....	202
5.4 Dreams as Low-Cost Counterweights	216
5.5 Research Notes: QS-Residuals After Nuisance Modeling.....	230
5.6 What Would Falsify QS?	248
Chapter 6 — Time Horizons and the Shadow Price	262
6.1 Why Endgame Balancing Intensifies.....	268
6.2 The Intuition of Shrinking Horizons	280
6.3 Hospice Across Cultures.....	287
6.4 Research Notes: Shadow Price λ_t and Horizon H_t	302
6.5 What to Measure (EEG/fMRI, Time Perspective)	315
6.6 Fail Patterns for Horizon Scaling	328
6.7 Population Shadow Price and Policy Windows	337
Part IV — Measuring Feeling Without Fooling Ourselves	339
Chapter 7 — The Hedonic Composite Index (HCI)	341
7.1 Inputs: Report, Physiology, Brain, Behavior, Dreams.....	344
7.2 Why Composite Beat Single Meters	356
7.3 Keeping It Honest: Blinds and Preregistration.....	361
7.4 Research Notes: Latent (CFA/IRT) and State-Space.....	369
7.5 Hedonic Composite Units (HCU)	375
7.6 Fail Conditions for HCI.....	379
Chapter 8 — “Same Scale” Across People and Places	383
8.1 The Invariance Problem.....	387
8.2 Culture and Age Effects.....	394

8.3 Universal Anchors (Pain, Chills, Social Exclusion)	405
8.4 Research Notes: Configural → Metric → Scalar Invariance.....	417
8.5 Calibration Ladder (Within → Between → Cross-Cultural).	428
8.6 Propagating Uncertainty into the Ledger	439
Part V — Identity and Edge Cases	449
Chapter 9 — Unity of the Stream	452
9.1 Conscious Access: One Stream or Two	456
9.2 The Unity Index (Plain Speech)	461
9.3 Pauses: Sleep, Anesthesia, Coma.....	463
9.4 Split-Brain and DID	474
9.5 AI and Brain Organoids.....	485
9.6 Research Notes: Blinded Adjudication and Thresholds.....	497
Part VI — Evidence We Can Look For (Right Now)	505
Chapter 10 — Dreams: The Night Workshop	510
10.1 What Dreams Do for the Ledger	514
10.2 Classic Observations	519
10.3 Predictions: Valence Inversion After Tough Days	523
10.4 Research Notes: REM Timing, Sampling, Coding.....	528
10.5 A One-Week Dream Ledger Exercise	536
10.6 Fail Patterns in Dream Data	541
Chapter 11 — End-of-Life: Where the Law Shows Its Hand.....	547
11.1 Why This Is the Sharpest Test.....	556
11.2 What Hospice Workers See	562
11.3 Ethics: What We Will and Will Not Do	572
11.4 Research Notes: Variance Compression and Neural Signatures	582
11.5 Reading Anecdotes: Scripts vs. Signals	596
11.6 Fail Patterns at Terminal Closure.....	607
Chapter 12 — The Lab Bench: Horizon Tasks and TMS	622
12.1 Short vs. Long Horizons in the Lab.....	631

12.2 Expected Control-Hub Signatures	634
12.3 Perturbation: TMS to rIFG/ACC.....	642
12.4 Research Notes: Preregistration, Power, ROIs	651
12.5 Negative Controls	661
12.6 Fail Patterns in Lab Tests	671
Chapter 13 — The Long View: Telemetry Across Years.....	678
13.1 Longitudinal HCI, Practically	691
13.2 Compression Near the End.....	699
13.3 Life Events as Ledger Shocks	710
13.4 Research Notes: Missingness and Hierarchical Models	722
13.5 Citizen Science Without the Creepiness	734
13.6 Fail Patterns: Expanding Variance	746
Part VII — Rival Explanations, Fairly Presented	756
Chapter 14 — The Hedonic Treadmill and Opponent Processes	761
14.1 Best Evidence For	763
14.2 Where They Shine	769
14.3 What They Cannot Guarantee	774
14.4 Research Notes: Tendency vs. Law	781
14.5 If Rivals Win, What LoF Learns	789
Chapter 15 — Predictive Coding and Free-Energy.....	796
15.1 The Big Idea (Uncertainty Minimization)	799
15.2 Affect as Prediction Error	802
15.3 Why Viability ≠ Fairness.....	809
15.4 Research Notes: Overlap vs. Independence Tests	816
15.5 If Rivals Win, What LoF Learns	830
Chapter 16 — Reinforcement Learning and Homeostasis	835
16.1 Rewards, Set Points, and Care	838
16.2 Optimization Isn't Balance	844
16.3 Composite Rivals (Hybrids)	851

16.4 Research Notes: Model Comparison and Adversarial Fits	858
16.5 If Rivals Win, What LoF Learns	867
Part VIII — Evolution and Simulated Worlds	874
Chapter 17 — Natural Selection Meets a Law.....	877
17.1 Constraints the Genome Can't Break	883
17.2 Why Control Systems Resemble QS	889
17.3 Cultural Echoes: Karma, Justice, Penance	895
17.4 Cross-Species Predictions	901
17.5 Research Notes: Fitness-Neutral, Constraint-Binding.....	909
17.6 Fail Patterns: Species with Systematic Imbalance.....	927
Chapter 18 — If Life Is a Game	935
18.1 Why a Designer Would Bake in LoF.....	940
18.2 Constraints Beat Patches (Cost and Elegance).....	942
18.3 Dreams as Offline Balancing Passes	946
18.4 Research Notes: No-Neutrality-by-Fiat in Code.....	952
18.5 Indirect Evidence: Worlds That Fail Without Constraints	959
18.6 Fail Patterns in Simulation Studies	964
Part IX — Ethics and Human Dignity	971
Chapter 19 — What This Never Justifies.....	974
19.1 No License to Ignore Pain	978
19.2 Duties of Caregivers and Researchers	983
19.3 Justice Aimed at Restoration.....	990
19.4 Research Notes: Non-Sentience in Simulation.....	997
19.5 Communication Ethics	1005
19.6 Hard Lines We Will Not Cross	1012
Chapter 20 — Hope, Freedom, and Daily Life	1020
20.1 Freedom Inside Guardrails	1024
20.2 Meaning Without Illusions	1029
20.3 If the Law Is True	1033

20.4 If the Law Is False	1039
20.5 Talking with Skeptics	1043
20.6 A One-Page Summary to Share	1048
Part X— Applying the Law: From Habits to Labs	1051
Chapter 21 — The Ledger Gym: Habits, Queue Traps and Repair	1054
21.1 Warm-Up: Your Life as a Ledger Gym.....	1059
21.2 Daily Habits as Guardrails and Repairs	1062
21.3 Seven Queue Traps (Classical “Sins” Reimagined).....	1068
21.4 Social Relationships: External Guardrails and Repair Kits	1074
21.5 Ritual and Reflection: Spiritual Practices as Queue Hygiene	1078
21.6 N-of-1 Experiment: Training Your Own Guardrails	1083
21.7 What Not to Do: Ethical and Practical Warnings	1088
Chapter 22 — The Scientific Playbook	1092
22.1 Full Prereg Packages (drop-in, decision-grade).....	1097
22.2 HCI Code and Open Data Hygiene (drop-in, LoF-native)	1103
22.3 Multi-Site Replication.....	1107
22.4 Red-Team Challenges and Bounties	1116
22.5 Research Notes: Equivalence Testing for L(T)	1125
22.6 Classroom Dream Counterweights	1133
22.7 A Simple Horizon Task	1136
22.8 Ethics and Blinds for Teens	1140
22.9 Research Notes: Consent Templates.....	1145
Part XI — The Case in One Place.....	1149
Chapter 23 — The Ten Hardest Objections (and Our Answers)	1153
23.1 “You Can’t Measure Feelings”.....	1158
23.2 “Adaptation Explains It”	1160
23.3 “You’re Moralizing Physics”.....	1163
23.4 “Identity Is Fuzzy”	1166
23.5 “Dreams Are Noise”	1170

23.6 “The Brain Is a Prediction Machine”	1173
23.7 “Simulations Prove Nothing”	1177
23.8 “Evolution Wouldn’t Select This”	1181
23.9 “It’s Unfalsifiable”	1186
23.10 “It’s Dangerous to Say Suffering Balances”	1190
23.11 Research Notes: Where to Find the Evidence	1194
 Chapter 24 — If Fairness Is Real.....	1197
24.1 The Discovery Claim	1204
24.2 What Ordinary People Can Do	1210
24.3 A Call for Courage (to Test, Not Believe)	1220
24.4 Utopia Thought Experiments: External vs. Internal Moderation	1227
24.5 Bridge to Synthesis	1234
 Chapter 25 — Final Synthesis	1242
25.1 The Ontological and Metaphysical Perspective	1247
25.2 The Physical and Systems Perspective	1251
25.3 The Psychological Perspective	1261
25.4 Spiritual and Moral Parallels	1273
25.5 Societal and Ethical Implications	1280
25.6 Final Reflections	1287
 Summary of Main Ideas	1292
Glossary of Key Terms	1317
Notation and Mathematical Conventions.....	1323
Core Formulas and Equations.....	1336
Citations and References	1343

Illustrated Guide to Fairness



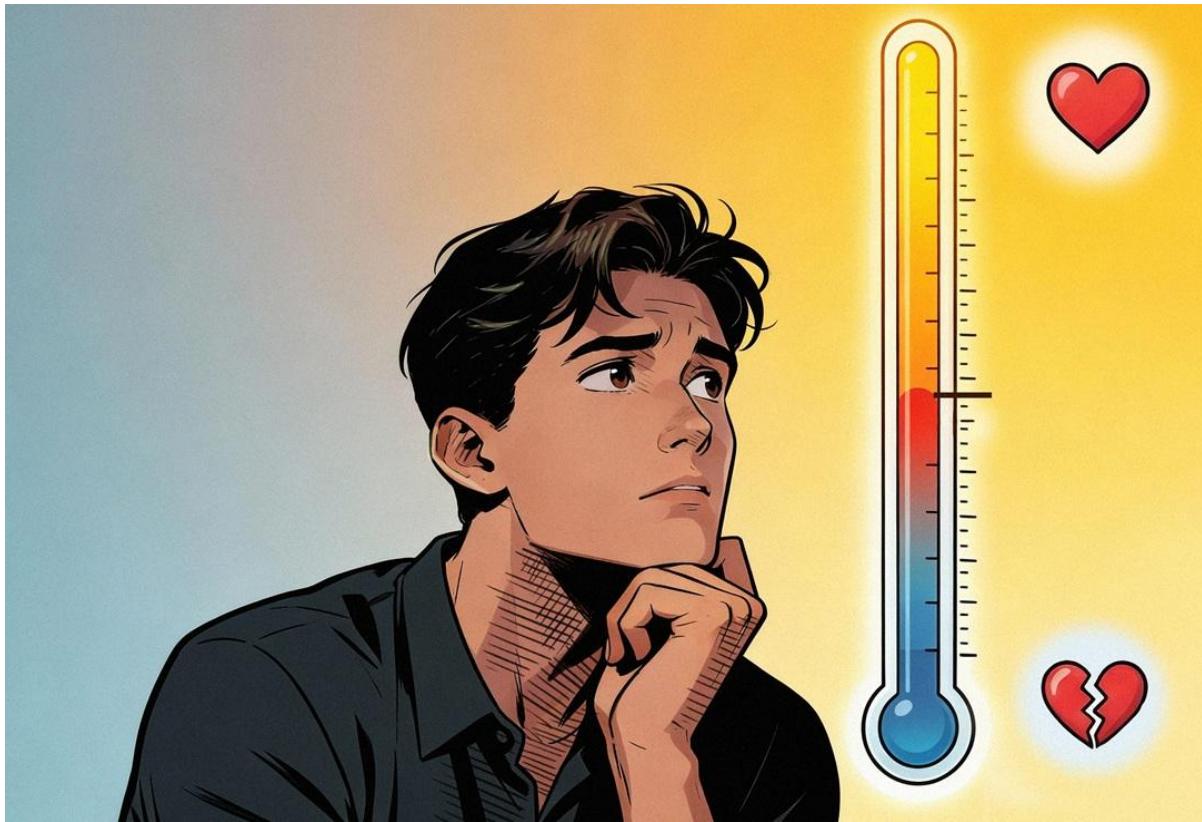
From a young age, we are told that “life is not fair.” As we look at the world, the evidence seems overwhelming. Some lives appear to overflow with unearned comfort and joy, while others are marked by relentless struggle, loss, and tragedy. On the surface, the distribution of pain and pleasure often looks unjust and even cruel. Facing this disparity, it is only natural to ask: Why is life so unfair?



However, appearances can be deceiving. We don't see the entire story of another person's life. A billionaire may inhabit a private hell of constant fear, paranoia, and stress, while a person with nothing might access consistent, genuine moments of peace and pleasure. Every life contains hidden burdens that weigh it down and unseen blessings that lift it up. Regardless of how a life looks from the outside, the internal experience is always a complex, shifting mix of good and bad.



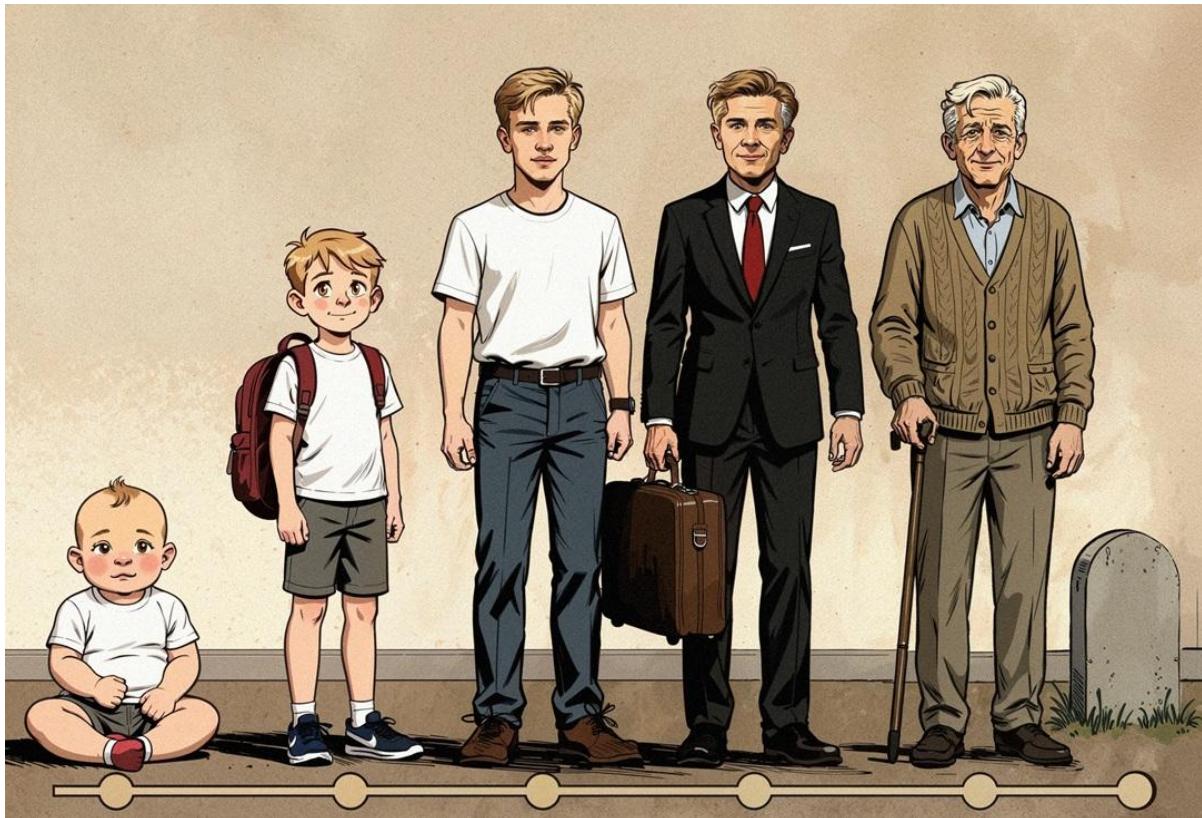
It is easy to judge someone else's life by how it appears, especially in a world where people carefully curate what they show to others. On social media, for instance, someone may post only their brightest, happiest moments, creating the illusion that their life is mostly joyful with less suffering. But these snapshots hide the complexity that every human life endures. Appearances might hide the late-night worries, the private losses, or the quiet frustrations that others carry. Because we rarely glimpse the full picture, it becomes natural to assume that other people have it better than we do and life itself must be unfair.



Even though life seems unfair at times, there is evidence that a deeper balance exists. A sudden stroke of luck rarely arrives without unseen costs. Pressure, fear, or loss often follow closely behind. And those who endure great hardship do not always remain broken. Many adapt, discover strength they didn't know they had, and often find joy in places they never expected. Psychologists have noticed this pattern and call it hedonic adaptation. It is as if our emotions carry a quiet thermostat, constantly pulling us back from extremes toward a balanced state.



There is one way life could be fair for everyone, no matter how different our lives appear. Not through money, status, or luck, but through *felt experience*. In this view, fairness would not depend on how long someone lives, how their life unfolds, or what they believe. It would depend only on what it feels like to exist. Perfect fairness would mean that, by the end of a conscious life, the total moments of pleasure and the total moments of pain balance one another. We call this hypothesis the Law of Fairness. This book explores whether such a law exists, and how it might shape every human life.



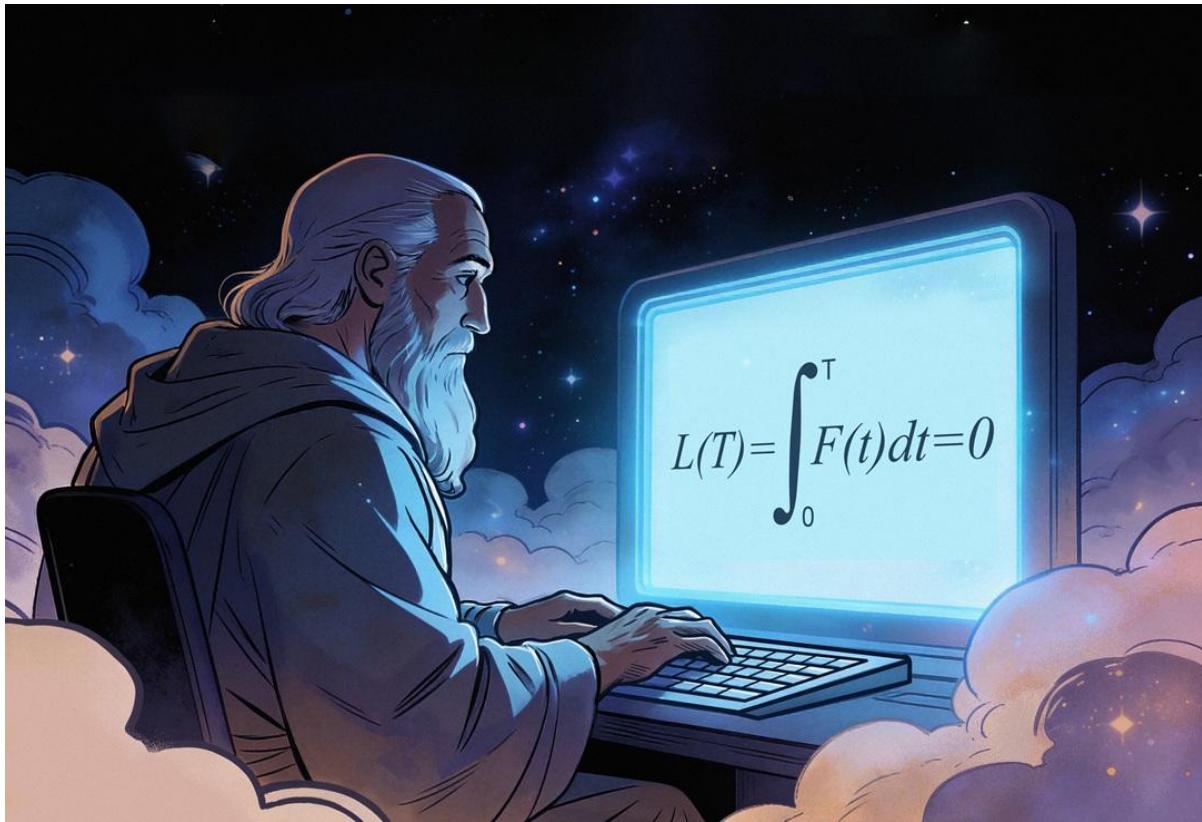
To determine whether life is truly fair, we must examine the entire story, not just a single chapter. Every experience—from the first moment of awareness to the final moment of consciousness—contributes to the overall balance. Joy and sorrow, pleasure and pain, success and failure all leave their mark on the total. Only by looking at a whole lifetime can we begin to see whether life, taken in full, is fair.



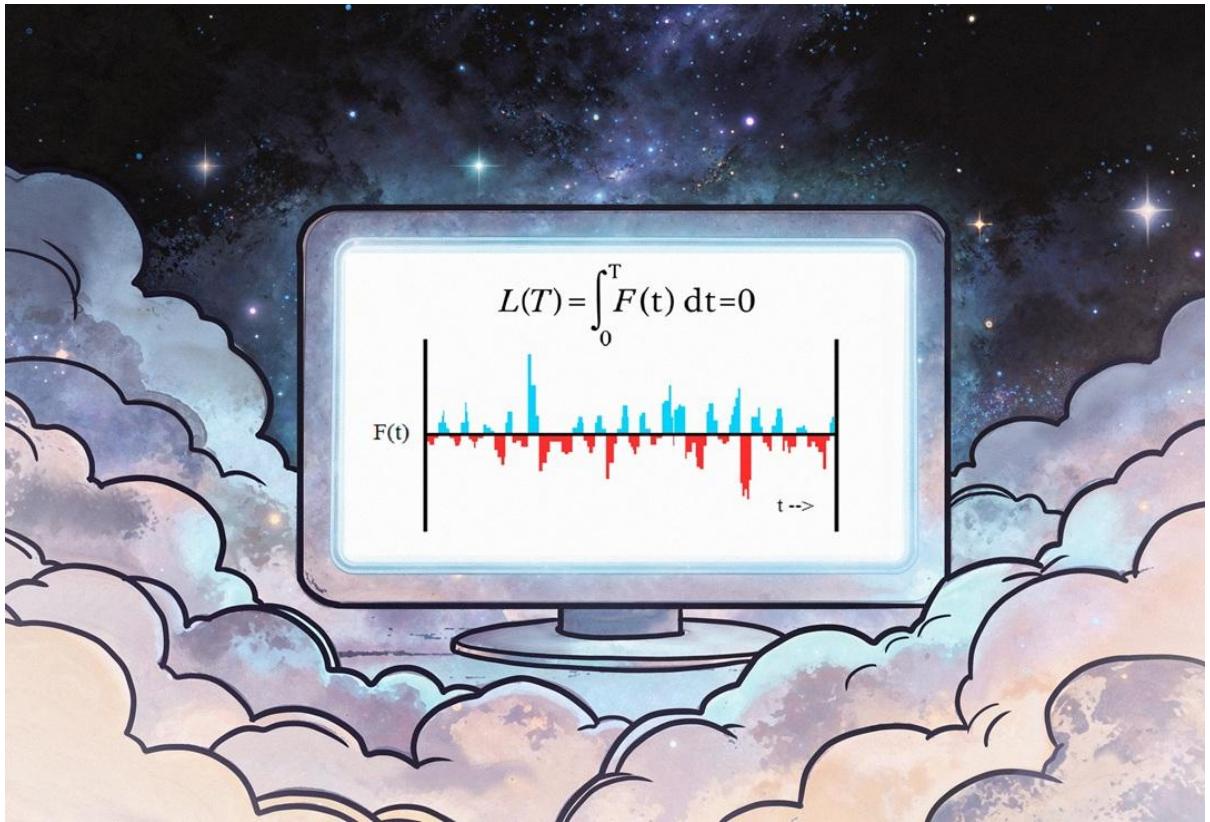
How does life keep things fair without us noticing? It works quietly through the currents of our own minds. Our instincts, desires, addictions, fears, and even sudden pulls of attention guide us toward experiences that help uphold the Law of Fairness. Urges that feel like free will often act as guardrails, steering us away from choices that would leave us permanently unbalanced. Under this law, a mind cannot “pass on” while it still owes a debt of feeling. Yet reaching equilibrium does not necessarily require death. Most of us find balance many times throughout our lives, daily or weekly for some people, and for others perhaps only once in a lifetime.



For every delight we enjoy, an equivalent cost in feeling must be paid. There is no “free” happiness. All the pleasures of life, food, play, rest, love, and companionship must be balanced by an equal measure of life’s pains, anger, sorrow, anxiety, and stress somewhere in the timeline, at some point in life. Yet this means none of our suffering is wasted. Each moment of pain or frustration functions as a credit, softening a future hardship or paying for a past joy. The Law of Fairness gives meaning to every struggle. Our tears today may water the flowers of tomorrow’s happiness or extinguish the lingering fires of the indulgent past.



If the Law of Fairness exists, then something must keep account of our feelings. We call this unseen process the Queue System. Rather than a single mechanism or controller, the Queue System is best understood as the emergent result of many interacting biological, psychological, and informational processes that regulate conscious experience over time. Taken together, these processes constrain the long-term accumulation of positive or negative felt experience across a unified conscious life. Imagine it as a program running beneath awareness, guiding experience toward balance. It ensures that no life ends owing joy or pain.



Consider a graph of a person's feelings over time. At each moment t , $F(t)$ represents the net felt value. Blue bars above the baseline show net positive felt experience, and red bars below represent net negative felt experience. The horizontal axis spans from the "Birth of Mind" on the left to the "Death of Mind" on the right. The Lifetime Ledger $L(T)$ is the running total of these values. A life may display wild swings, yet by the final moment, the total area above the line must equal the total area below it. This is the Law of Fairness in action: a conservation of feeling. Fairness is not measured in the world of things, but in the internal world of felt experience alone.



The Law of Fairness explains balance without invoking miracles. It proposes that fairness is a natural law, as fundamental as gravity or the conservation of energy. The Queue System acts as an invisible guardrail, shaping life toward balance rather than allowing a permanently unfair ending. We have free will, but only within the choices that preserve the ultimate balance. This limitation is an unseen filter on our decisions. Fairness emerges from the physics of experience itself.



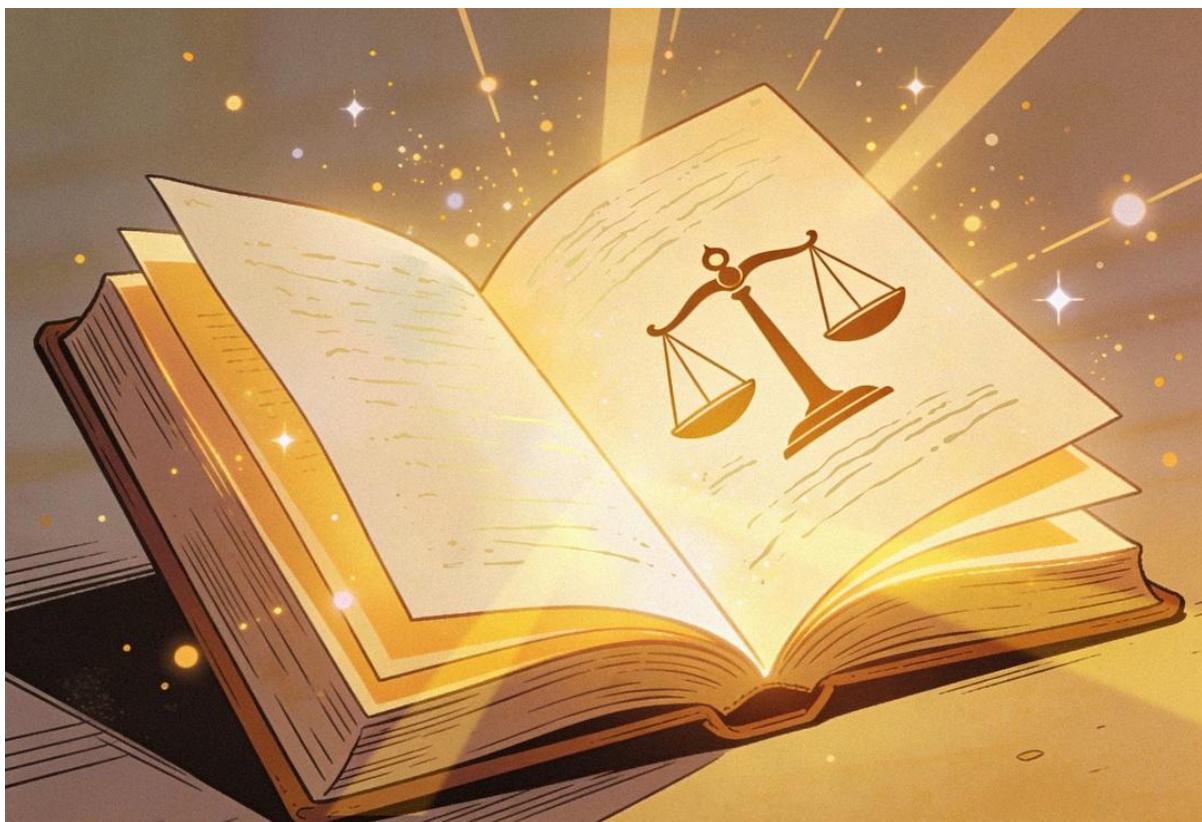
We cannot cheat the system. There is no way to take more than our share of good times without eventually accounting for it. The Queue System shapes our thoughts quietly, preventing paths that would leave a life permanently unbalanced. Certain ideas never take hold; certain choices feel impossible to sustain. Each mind moves within a shared field of consequence, where one person's excess or hardship can subtly affect others. Even dreams play a role, acting as a quiet workshop for emotion, allowing the mind to rehearse fear or joy to fine-tune the balance without altering reality.



In a universe governed by the Law of Fairness, not every outcome is possible. Any path that would lead to a permanently unfair result is filtered out before it happens. It is as if the universe ensures that no one ultimately ends with an irredeemably unfair lot. We choose, but only among options that can be balanced by the death of mind. Life is more constrained than it appears, unfolding along fewer possible trajectories than we imagine, because only fair endings are permitted to exist.



One way to view the Law of Fairness is through a spiritual lens. In this view, the Creator designed life to be fair not through constant intervention, but through structure. Fairness is automated, woven into reality like a natural law. For any afterlife, this framework suggests that each being would die with their emotional ledger complete. We do not carry excess joy or unresolved sorrow forward. No life is destined for endless pain without resolution, and no debt of feeling is left unpaid.



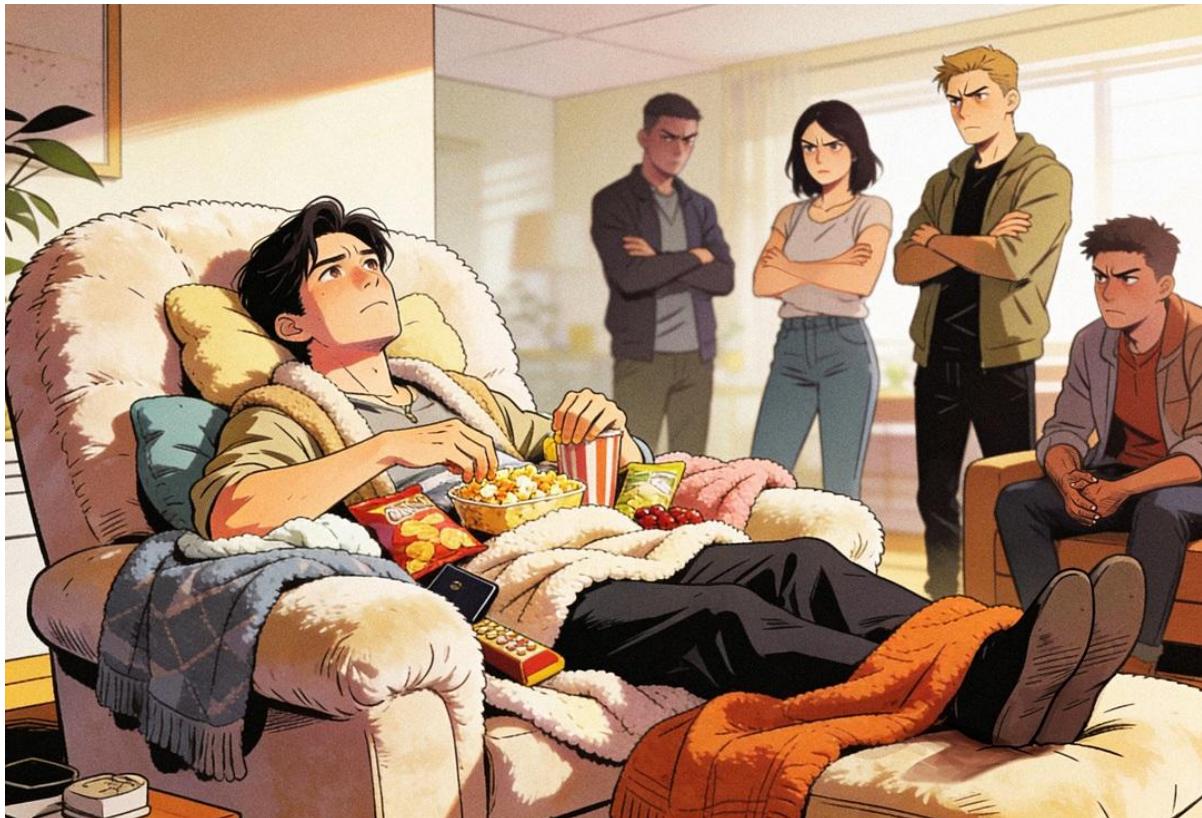
Across history, traditions offered guidance on how to live, how to treat others, and how to endure suffering. Long before we understood the Law of Fairness, stories, rituals, and moral frameworks helped people stabilize their lives and make sense of hardship. They preserved order and meaning in a world that felt unjust. We are drawn to beliefs and practices that help us carry joy and sorrow in proportion because we intuitively sense the ledger behind experience. In the end, we live by our balance, and we die by our balance.



Many have sensed a “missing piece” in our understanding of existence. The Law of Fairness might be that piece, the key connecting what science observes with what spirituality teaches, as well as what countless people have experienced firsthand. Together, these perspectives form one complete picture of reality. Scientists, philosophers, and seekers have pointed toward this for centuries, calling it hedonic adaptation, fate, karma, or divine will. The Law of Fairness may be the unifying thread that ties them all together, grounded not in wishful thinking, but in the nature of mind.



This perspective sheds new light on old ideas. Consider the “deadly sins.” Excesses like greed, gluttony, or pride are deadly because they flood a life with extreme pleasure that destabilizes the ledger. They push life so far out of experiential balance that the Queue System’s guardrails must intervene forcefully, compressing the trajectory to preserve fairness. If we live closer to balance, we reduce the need for these harsh adjustments. How much of the pain and physiological change that we associate with aging arises from repeated fairness correction?



Each of us shapes our own heaven or hell, in this life, by how we live. When we chase easy pleasure and avoid pain, imbalance accumulates. Over time, life introduces resistance—strained relationships, conflict, or friction—to restore equilibrium. Others may respond harshly to us, not out of malice, but as part of this balancing process. When we live with moderation, fewer corrections are required. No virtue exempts us from balance, and no fortune goes unpaid. Every pleasure carries weight, and every hardship contributes to the final accounting. Bad things still happen to good people so that life remains fair for everyone.



We pursue goals expecting profound satisfaction. We tend to think effort is the price of the goal, but the Law of Fairness reframes this: the strain of effort helps pay for the joy of achievement. Satisfaction must be earned somewhere in time. When effort is avoided, unearned pleasure pulls us out of alignment, often returning as failure or setbacks later. Some accomplishments demand hardship first because their fulfillment would otherwise push the felt experience too far into debt. Obstacles are not punishment but a necessary structure of fairness.



Think of your daily habits as entries on a balance sheet of pleasure and pain. Are there unnecessary pleasures you can trim? Mindless scrolling, extra indulgences, or idle comforts may feel good now, but they carry hidden costs that add up. Cutting back on easy comforts and instead choosing challenge is one way of managing the ledger. Focus on working hard, building good habits, and letting go of what doesn't matter. Discipline is not just productivity; it is an investment in future satisfaction, reducing the need for involuntary corrections later.



Whenever life feels unfair, remember the Law of Fairness. If you are enduring a hard time or have not accomplished a goal, recognize this difficulty as part of how balance is maintained. When something wonderful happens, cherish it, but do not overindulge. Trusting in this balance can bring comfort during struggle and humility during success. Every experience, whether welcome or painful, plays a role in guiding a life toward its fair conclusion. We are being steered toward neutrality, one experience at a time.



We are all part of one living fabric. Every act of care or cruelty sends ripples through the whole, shaping not only our own balance but the balance of those around us. While the Queue System guides experience toward balance, it never justifies causing harm or withholding kindness. Our duty is to ease suffering and lift one another, letting compassion protect the world's stability. Just as excess can destroy a single life, cruelty or indifference can destroy a society. Through moderation and care, we can live in ways that never threaten the Law of Fairness, allowing humanity's full potential to unfold.

Part I—The Question That Won’t Go Away

Some questions fade with age; this one refuses to. It confronts us everywhere: in hospital rooms where a child fights for life against all odds, in courtrooms where outcomes hinge on luck as much as law, in quiet kitchens late at night when we replay the day’s injustices. We feel it when life goes astonishingly well for someone who “did nothing to deserve it,” and when life goes horribly wrong for someone who seemingly did everything right. We accept that the weather is unfair, that markets are unfair, even that dice are unfair. But can a life be fair?

“Fairness” is easy to chant in slogans and hard to pin down in detail. We learn as kids that fairness means equal slices of cake, but adulthood teaches us a harsher lesson: the world often hands out slices with reckless abandon. Procedures can be fair, the same rules for everyone. Distributions can be fair, equal shares for all, or extra help to those who need it most. Yet at the end of a life, none of those surface arrangements guarantee what matters most: how it felt to be that person. If fairness is to mean anything at the deepest level, it cannot stop at rules or resources. It has to reach what every creature ultimately lives inside: the stream of experiences that make up a life.

This Part introduces the central idea of the book in everyday terms. You’ll meet three brief portraits, a child, a rockstar, and a monk, not to prove anything decisively, but to make the stakes of the question viscerally clear. You’ll see why popular answers (“life tends toward fairness,” “good people are happier,” “karma gets everyone eventually”) comfort us, yet still leave the hardest cases unsolved. And you’ll encounter the one-sentence claim we will test throughout this book: that each unified conscious stream ends with a neutral ledger of felt experience at the Death of Mind. Here, “unified conscious stream” refers to any system that genuinely supports integrated conscious experience; if such a stream exists, the Law claims it closes neutral at the Death of Mind. We do not offer this as a moral wish or a metaphysical promise, but as a constraint, a rule of admissible histories in a world like ours. In plainer terms, it’s like a proposed conservation law for conscious experience: impersonal and natural, not mystical or moralistic. And like any scientific hypothesis, it will stand or fall by the evidence. If real lives can end with a clear surplus of pleasure or of pain, then the idea in this book is wrong and should be set aside. Science deserves that clarity: under the Law of Fairness, the system must eventually balance out each life, not merely tend to on average.

What does a “neutral ledger” mean, without diving into equations? Imagine a transparent cylinder representing one whole life. Every pleasant moment drops in as a green bead, and every unpleasant moment as a purple bead. Over time, both the total number of beads and the counts of each color only ever increase. What changes is the ratio

between them. A single day can add far more of one color than the other, and entire decades can be skewed, even though no experiences are removed and nothing disappears. To make this picture precise, imagine that each bead is the smallest countable unit of feeling, already combining intensity and duration so that one pleasant unit offsets one unpleasant unit at the same granularity. Real experience is usually mixed, so a single interval contains a blend of both colors that simply accumulates over time. For example, a 90-minute period of mostly happy experience might, at fine resolution, be about 80% pleasant and 20% unpleasant; if that interval contained 5,000 total units, the ledger would record 4,000 pleasant units and 1,000 unpleasant units, for a net of 3,000. Likewise, a 30-second burst of near-bliss could be 99.7% pleasant and 0.3% unpleasant; if that interval contained 1,000 total units, the ledger would record 997 and 3, respectively, for a net of 994. The exact numbers are illustrative; the point is how the counting works. Across a whole life, all such mixed intervals are summed, and the Law of Fairness claims that, by the time the stream ends, no life is unending woe without relief and no life an endless parade of joy without answer. Periods of surplus on one side are, before the Death of Mind, counterbalanced by contributions on the other, so that when the ledger closes, the books are level.

If real lives can end with a clear surplus of pleasure or of pain, then the idea in this book is wrong, full stop. We set the bar that high because a law of nature should invite decisive tests. Under the Law of Fairness, balance must always occur eventually for each life, not merely “usually” or “on average.”

You won’t need any new jargon to read this Part. When a technical term appears later on (for example, the hedonic composite index that quantifies feelings, the Queue System (QS) that models how life might “queue up” compensating events, or horizon scaling as one’s remaining time shrinks), we’ll introduce it in plain language first and then, when needed, unpack the details in a Research Note. You can ignore those notes and still follow the story, or focus only on the notes and still see the logic made explicit. Both tracks convey the same truth in different voices.

Why begin with stories? Because everyone already carries a private dataset of one: your own life, plus the lives you’ve observed up close. You’ve seen rebounds after grief, and cruel downturns that didn’t rebound. You’ve seen last-minute peace and last-minute panic. You’ve seen people change for the better, and people stay inexplicably the same. A good theory of fairness must be large enough to hold all of that without explaining it away. It must never ask you to pretend that someone’s suffering was “actually fine” just because a graph later balances out. In this Part, we set a tone that will never change: balance at the end is not an excuse for pain in the middle. Present pain deserves present

care. If balance is real, it happens through human choices, through rescue, repair, apology, art, rest, medicine, not around them or in spite of them.

What this Part will do for you:

- Name the target. By “fairness,” we mean something that holds for every conscious life, not merely on average and not only for the lucky few. No exceptions.
- Clear away near-misses. We’ll show why equal rules, equal resources, good intentions, or the idea that “most people bounce back eventually” are not sufficient answers to life’s most unfair moments. These well-meaning ideas each miss the mark in crucial ways.
- State the claim cleanly. The Law of Fairness is presented as a conservation-style constraint on lifetime felt experience. If the law is true, the system must leave certain telltale signatures in the data; if it’s false, there are specific patterns that will eventually prove it false. If a claim can never be falsified, it isn’t worth much.
- Set expectations. There will be no metaphysical guarantees and no hand-waving sermons here, and we won’t soften the sharp edges of reality. Expect careful language and testable ideas, hypotheses you can hold to evidence, rather than comforting platitudes. This is about uncovering a law of nature, not affirming what we wish were true.

Chapters in this Part:

- **Chapter 1 — Why Fairness? Why Now?** – Confronts the age-old question of fairness head-on and asks why this persistent problem demands a fresh, scientific look right now. Chapter 1 lays out the problem in full daylight and frames “fairness” as a literal, testable rule rather than a vague ideal.
- **Chapter 2 — Feelings As the Final Currency** – Argues that the true currency of fairness is felt experience. In Chapter 2 we explain why subjective feelings (pleasure, pain, and the full spectrum between) matter more than any external scorecard of wealth or luck. We also sketch how feelings might be measured and why any metric must respect human dignity and cross-cultural consistency.

Where we go next:

We turn to Chapter 1, where the question is put on solid ground. First, three portraits (a child, a rockstar, a monk) fix the target at the life-level stream of experience. Next, we sort everyday meanings of “fairness” and show why none of them guarantees a fair life in felt experience. We then face the hard cases that break easy theories, and explain why a mere tendency toward balance is not enough. The chapter closes by outlining how to read the book, so the argument is rigorous without losing the human thread.

Chapter 1 — Why Fairness? Why Now?

Late one night in a pediatric hospital, a parent asks why her terminally ill child must suffer while others lead easy lives. Across the country, a famous rockstar trashes a hotel room in a drug-fueled rage, then wakes up unharmed and carefree, once again escaping consequences that would crush most people. And in a remote monastery, a monk who has lived a life of compassion and sacrifice struggles with a painful, lingering illness that no prayer can heal. Three very different scenes, one identical question: Where is the fairness? How can a world subject a kind child or a saintly monk to agony, yet let an irresponsible celebrity skate through life untouched? We've all felt that private pang, the quiet anger or confusion at how unfair life seems sometimes. Fairness isn't just a child's complaint; it's a ghost that follows us into adulthood, appearing whenever life's scales seem grossly out of balance.

"Fairness" is an old word with a hard job. It stands between what we hoped life would be and what it often is. As children, we start with the simple idea that fairness means sharing equally and playing by the same rules. By adulthood, we've watched luck and injustice routinely trump those rules. Some of us encounter fairness in the dry language of courtroom procedure: did everyone get a fair hearing? Some encounter it as a distant promise in religious or spiritual teachings: someday, somehow, all wrongs are righted. But for most of us, fairness is something far more personal and visceral: it's the word we reach for when someone we love is hurt for no good reason. It's the ache when a life goes sideways despite everything done right. The central question of this book is whether fairness can be literal, not just a comforting slogan, not just moral window-dressing, but an actual rule that governs how lives actually go. Can fairness be more than an ideal? Could it be a law of nature hiding in plain sight?

To ask that question in the twenty-first century is to ask it with new tools and new responsibilities. In earlier eras, people could only wonder or take it on faith that "everything happens for a reason." Today, we don't have to simply accept or reject that hope; we can measure things we never could before. We can now monitor brain and body signals with noninvasive sensors, even outside the lab and over long periods, capturing objective clues about pleasure, pain, stress, and comfort in daily life. We can model hidden patterns, stitching together many subtle signals into a single, more reliable index of someone's lived experience. We can preregister predictions about life trajectories and then let the data embarrass us if we're wrong. In short, we can finally treat fairness not just as a philosophy or wish, but as a hypothesis to be tested. And we can do all of this while keeping human dignity at the center, especially when dealing with illness, suffering, and end-of-life contexts, because ethics is not decoration; it's the first method. Put

another way: relief is a systems variable; comfort and dignity override data collection. We will never pursue knowledge at the expense of compassion.

We also live in a time of frayed trust, which makes it all the more important to proceed carefully. Many people feel that the “rules” of society, and by extension, the rules of life, mostly favor the already lucky, and they are understandably cynical of grand notions of fairness. Others worry that meaning itself is dissolving under a tide of statistics, that by reducing everything to data we risk losing the human story. A book about fairness runs on twin risks: it could preach feel-good morals and offer false comfort, or it could retreat into clever mathematical theories that never touch the reality of a hospital bed or a funeral. We will do neither. Our standard is simple and stark: if fairness is real in the way that matters most, it must show up in the actual experience of living, by the end of a life. If it is not real, if life can truly deal out irredeemable injustice, then the idea should be set aside. We won’t cling to a comforting fiction. In science, a theory earns its keep by allowing itself to lose. The Law of Fairness will earn its keep by making clear predictions that could absolutely fail. If the “rules of life” as we define them mostly favor the lucky, or if our analyses end up dissolving meaning rather than illuminating it, we’ll know we took a wrong turn.

So what exactly are we proposing to test? We can state it in one sentence: Every unified conscious stream ends with a neutral ledger of felt experience at the Death of Mind. That is the claim, in plain language, that we will examine from every angle. If even one unified conscious stream can be shown to reach the Death of Mind with clear residual imbalance beyond the defined neutrality band, the Law fails. In essence, the Law of Fairness says that no matter how wildly life’s day-to-day or decade-to-decade experiences fluctuate, by the time an individual conscious life comes to an end, the total sum of its felt pleasures and pains will balance out to neutral. It’s a bold claim. To even entertain it, we will need to define each part of that sentence rigorously: What counts as a “unified conscious stream”? What do we mean by “felt experience,” and how can we tally it? What constitutes the Death of Mind, exactly? The chapters ahead (and indeed the whole book) exist to pin down those definitions, to show how such a balancing mechanism could possibly be real without invoking magic or cosmic justice, and to lay out exactly what evidence would falsify the claim. We will name specific patterns, measurable, observable patterns, that must be present if the Law holds. And we will be equally clear about patterns that, if observed, would break the Law. If real lives can end with a giant surplus of unearned joy or irreparable sorrow, then the Law of Fairness is wrong, full stop, and our investigation will point to that failure. Clarity is our compass: a law that can never be proven wrong isn’t really a law at all.

In the rest of this chapter, we unpack why this question matters and set the stage for testing that one-sentence claim. We'll start by confronting the intuitive evidence around us, the stories and cases that make people doubt fairness in the first place, and then we'll clarify what we mean (and don't mean) by fairness. By the end of Chapter 1, we will have a precise statement of the problem and a roadmap for how to investigate it. Here's what you can expect to gain from this chapter:

What you'll get from this Chapter:

- A deeper understanding of why the question of fairness refuses to die out, and why it's more than just a childish complaint; it's a profound challenge that every generation grapples with.
- Clear insight into why the usual answers fall short: we'll see why common ideas like "everything happens for a reason," "good people will be happier," or "life tends to even out eventually" don't truly solve the hardest cases of injustice.
- A concise formulation of the Law of Fairness hypothesis in everyday language, a testable claim about life outcomes that we can either substantiate or refute with real data.
- An appreciation of how new science and ethical research methods can tackle this problem today: you'll learn why now is the time to treat fairness as a hypothesis (thanks to advances in measuring experience), and how we'll safeguard human dignity in the process.
- A preview of how this book will proceed, including the roles of human stories, measurable metrics, and "fail-safe" criteria, so you know what to expect in the chapters to come.

Subsections in this Chapter:

- **A Child, a Rockstar, a Monk** – Three vivid life portraits that illustrate, in turn, tragic misfortune, absurd good fortune, and the search for meaning. These cases put concrete faces on our core fairness question.
- **1.2 What People Usually Mean by Fairness** – A look at the common notions of fairness (from equal opportunity to cosmic justice) and how each captures only a slice of what we really crave when we say "that's not fair."
- **1.3 The Hard Cases No One Likes** – An honest confrontation with the cases that make even optimists squirm: innocent suffering, undeserved luck, and last-moment downturns that challenge any notion of balance.
- **1.4 Why "Tendency Toward Fairness" Isn't Enough** – Analysis of the comforting idea that "in the long run, things tend to balance out." We'll see empirically and logically why a mere tendency is too flimsy, why only a true law of fairness would be meaningful.

- **1.5 How This Book Works** – A brief guide to the structure of the argument and the research to come. This section explains our approach (blending narrative with analysis), introduces tools and terms we'll rely on, and sets you up for the journey into the next chapter and beyond.

Where we go next:

We begin, fittingly, with real lives in real predicaments. In Section 1.1, we'll meet our child, our rockstar, and our monk up close. Each of these three lives will shine a light on a different facet of the fairness question, grounding our discussion in concrete reality. Let's step into their stories and see what they reveal about the problem we're aiming to solve.

1.1 A Child, a Rockstar, a Monk

Stories don't prove a law, but they tell us what any honest law must be able to hold. Here are three lives sketched without sentiment. Each asks the same question in a different voice: What would fairness have to mean here?

The Child. A seven-year-old girl spends more time in the hospital than at home. Some mornings are ordinary, cartoons, a bite of toast, a sibling's joke. Many are not, needles, nausea, a parent juggling leave forms and fear. She is brave in the way children are: she sits still because the nurse asks. She does not bargain with the universe. She doesn't "deserve" her illness, no child does, and yet no medicine can give the total relief she needs.

If fairness is about equal procedure, the hospital does well, consent forms, careful dosing, infection control. If it's about equal resources, charities help with travel and meals. Yet none of those equalities touches the question that keeps her mother awake: In any sense, can my daughter's life be fair? Not someone else's life or tomorrow's chances, her stream. Her ledger.

A theory that stops at rules or resources cannot answer that mother. If fairness is real, it must speak to what it was like to be that child, summed across all her days. The Law of Fairness, if true, speaks at that level or not at all.

The Rockstar. A guitarist wakes to applause more often than alarms. He is gifted and charming and has never filled out a tax form without help. People give him things because it feels good to be near the light. There are real costs, jet lag, paparazzi, constant pressure, but the life is undeniably gilded. He isn't cruel. He's also no saint. Luck and talent did most of the lifting.

If fairness is desert, you get what you earn, his surplus raises doubts. Does one catchy melody deserve ten lifetimes of comfort? If fairness is utility, maximize total happiness, the stadium euphoria might justify it. But the question that follows him into quiet rooms is different: Can someone ride a tide of good feeling all the way to the end while others sink? If yes, then some ledgers close with surplus and others with debt. If no, if a conservation law governs streams, then even his life must balance out. Not by punishment, not by a last-act tragedy, but by a quieter arithmetic that only counts what he truly felt.

A notion of fairness that can only scold or fantasize about the rockstar isn't the fairness we need. The Law of Fairness claims that even a lucky life is a balanced life by the time the stream ends.

The Monk. A man retreats to a monastery because the noise in his head was louder than traffic. He learns to sit. He learns to notice a swallow of tea, a cedar scent after rain, anger arriving at the gate and leaving unfed. He gives away the career he once adored and finds, to his surprise, that his days are not smaller but more precise. He is not always serene. He sometimes misses fast internet and a certain person's laugh. He keeps his vows, stumbles, then keeps going.

If fairness is preference satisfaction, he looks foolish, he declined what many people crave. If fairness is happiness level, he often scores high, but in a way that doesn't depend on bending the world to his will. His life raises a subtle point: balance is not the same as blandness. He still feels grief when a parent dies, pride when a novice succeeds, shame when he speaks too sharply. The difference is that his available thoughts, what comes to mind, what can be chosen, have shifted. He can accept and repair more readily. In this book's terms, the Queue System may have thinned certain options and strengthened others, making emotional compensation more reachable without grand drama.

A notion of fairness that requires identical lives would erase the monk on sight. The Law of Fairness requires no sameness. It claims only that each stream, however loud or quiet or wild its path, closes neutral. Why these three? Together they corner the problem. The child forces us to ask whether fairness can touch the most undeserved harms. The rockstar forces us to ask whether fairness can coexist with extreme good fortune. The monk forces us to ask whether fairness allows real difference without smuggling in moral judgment. If a theory dodges any of these, it isn't about life-level fairness. If it faces all three, it must talk about felt experience across a lifetime, and nothing less.

1.1.1 Where we go next:

We have met the child, the rockstar, and the monk to fix the target of our inquiry where it belongs: inside the lived stream of experience. Next, we will sort the many everyday meanings of "fairness," rules, resources, opportunity, desert, aggregates, and place them where they help without letting any of them stand in for what it was like to be a life. In 1.2 we will name each common meaning clearly so that, from this point on, "fairness" in this book cannot be mistaken for anything but a life-level claim about felt experience. In the chapters ahead, we won't return to these people as case studies to be neatly "balanced out," that would be disrespectful and simplistic. We return to them as tests of our language. Whatever we claim must make sense in their company. If the Law of Fairness cannot speak in a hospital room, on a tour bus, and in a cloister without changing tone, it isn't yet a law worth testing.

1.2 What People Usually Mean by Fairness

When people say “that’s not fair,” they rarely all mean the same thing. Fairness does many jobs. It can mean same rules, same shares, you got what you deserved, the worst-off got help, or simply “the outcome felt right.” Before we ask whether a life can be fair, we need to unpack these familiar meanings, and see why, on their own, none of them can guarantee fairness at the level that matters most: how a life felt as a whole.

1.2.1 Procedural fairness: the rules were the same

Courts, contests, and classrooms rely on procedural fairness, equal rules, due process, impartial referees, published criteria. It’s the kind of fairness we can see and audit. When it’s missing, we protest with reason.

But procedure is a wrapper, not a life. The seven-year-old in the hospital may receive flawlessly fair procedures; the rockstar may face the same tax rules as everyone else; the monk may be treated even-handedly by his order. None of that touches their ledgers of felt experience. The procedure can be immaculate and the life still lopsided. Procedural fairness is essential for trust, but by itself it doesn’t make a life fair.

Takeaway: Rules can be fair while outcomes are cruel or lucky. Good procedures prevent some injustices; they do not ensure a fair life-level balance.

1.2.2 Distributive fairness: the shares were equal (or need-based)

Another common meaning is distributive fairness, equal slices for all, or more for those who need it most. This matters for dignity and survival. Just distribution also prevents many harms.

Yet equal resources don’t translate into equal felt outcomes. Two people with identical incomes can live very different emotional lives, because their bodies, histories, and contexts convert resources into experience differently. A need-based policy can still leave one person with months of chemo and another with no illness at all, no one’s fault, but someone’s life. Distribution is a lever on conditions; it isn’t a ledger of experiences.

Takeaway: Fair inputs are not the same as fair experiences. Better distribution helps, but it cannot guarantee a neutral closure for each life’s ledger.

1.2.3 Equality of opportunity: the race started at the same line

“Level the starting line,” we say, remove barriers of class, race, gender, money. This is morally urgent and empirically valuable; opportunity correlates with health and happiness.

But opportunities are probabilities, not experiences. Even in the fairest meritocracy imaginable, sickness, accident, and sheer timing still intervene. The child dealt a rare genetic disorder did not lack opportunity; the ledger of her life is still hers. Equal starts cannot ensure that every stream ends balanced.

Takeaway: Opportunity is about chances; life-level fairness is about what actually happened and how it felt in the end.

1.2.4 Desert: you got what you earned

We admire effort and skill; we recoil at freeloading. Desert allocates rewards by contribution. It motivates effort and seems commonsensical.

Yet what any of us manages to “earn” is saturated with moral luck, genes, parents, mentors, historical moment. The rockstar’s gift and timing are not faults, but they are not deserts in any pure sense either. More importantly, desert-based fairness struggles with cases of non-agency: infants, coma patients, people with advanced dementia. They feel, but they aren’t “earning” anything in a market sense. A standard of fairness tied only to desert cannot cover the very cases that most cry out for fairness.

Takeaway: Desert can guide who deserves rewards in a society; it cannot ensure just ledgers for those who suffer without agency or reward those who experience great fortune unbidden.

1.2.5 Consequentialist sums: the most happiness for the most people

Utilitarian and prioritarian ideas ask us to maximize total (or weighted) happiness, the greatest good for the greatest number, with maybe extra weight for the worst-off. These are powerful for policy because they make trade-offs explicit.

They are also notorious for sacrificing individuals to aggregates. A world that adds a tiny bit of bliss to a billion people by subtracting huge happiness from one unlucky person “wins” on paper while destroying a life. Aggregates ignore the unit of fairness that counts: the conscious stream. Any standard that says “overall it worked out” while a particular life ends deeply negative is not a standard of life-level fairness.

Takeaway: Aggregate welfare is useful for policy optimization; it’s not sufficient to claim that each life closes balanced.

1.2.6 “Happiness science”: most people adapt

Psychology offers comforting statistics: hedonic adaptation (the “hedonic treadmill”), opponent processes, set-point theory. Indeed, many people do tend to drift back toward baseline after gains and losses.

Two cautions. First, these are tendencies, not guarantees; chronic pain, treatment-resistant depression, protracted trauma, and late-life collapse exist. Second, even if adaptation holds on average, it says little about terminal neutrality, what the ledger reads at the Death of Mind. Averages across time slices or across people do not speak for this stream, this ending.

Takeaway: Adaptation explains rebounds; it does not establish a law forbidding unfair endings.

1.2.7 Karma and afterlife promises

Many traditions place balance elsewhere: another life, heaven, final judgment. These stories guide communities, but they are not testable by our methods. The question here is whether a this-life balance can be stated and tested without metaphysical escape hatches.

Takeaway: Spiritual accounts of ultimate justice are meaningful to many, but outside empirical scope. This book tests a here-and-now claim.

1.2.8 Restorative justice and repair

Restorative approaches aim to heal rather than punish: truth-telling, apology, restitution, community work. They matter because relationships, not just rules, make lives livable.

Still, repair is local and partial. It can reduce suffering going forward. It cannot always repay what was already felt, nor guarantee the final balance of a particular stream. If the Law of Fairness is true, restorative acts may be major channels through which compensation occurs; if it's false, we still pursue them because they help in the present. Either way, their existence doesn't prove a life will end fair.

Psychologically, many people hold a "just-world" belief that everyone gets what they deserve in life (Lerner, 1980). This cognitive bias speaks to our deep desire for balance, but it is termed a "fundamental delusion" for a reason; assuming cosmic fairness can lead to complacency or victim-blaming without evidence. We must separate such comforting beliefs from any law to be tested.

Takeaway: Restoration and repair are essential practices, but they aren't proof of life-level fairness. They help now; they don't ensure a zeroed-out ledger at the end.

1.2.9 A working definition for this book

Having surveyed the usual meanings, we can now say what we will mean by fairness in the rest of this book:

Fairness refers to a property of a unified conscious stream such that the lifetime ledger of felt experience (pleasant minus unpleasant, properly measured and uncertainty-bounded) reaches neutrality at the Death of Mind.

This definition doesn't dismiss procedures, distributions, opportunities, desert, or repair; it relocates them. They are means by which better days are made and (if the Law holds true) channels through which balance is realized. But they are not the criterion. The criterion is simply what it was like, in total, to be that stream.

1.2.10 How we will keep the meanings straight

To avoid talking past each other, this book will use the following language precisely:

- Procedural fairness: same rules, transparent processes.
- Distributive fairness: how resources or outcomes are allocated.
- Opportunity fairness: barriers removed at the start.
- Desert: rewards tied to contribution or effort.
- Aggregate welfare: population totals or weighted sums.
- Life-level fairness (our target): neutral closure of a stream's hedonic ledger.

Whenever we claim evidence for or against fairness in this book, we mean life-level fairness – not merely the other kinds (valuable as they are).

1.2.11 Back to our three lives

For the child: procedural and distributive fairness are necessary kindnesses, but they don't tell us whether her stream can end balanced.

For the rockstar: desert and aggregate utility can rationalize his surplus, but they don't answer whether his ledger truly closes at zero.

For the monk: high subjective well-being doesn't imply bland sameness or moral superiority – it suggests his available thoughts make compensation easier without spectacle (QS quietly at work).

1.2.12 Where we go next:

The question that remains is the one this book exists to face: Can balance at the end be more than a metaphor? The next section (1.3, The Hard Cases No One Likes) lays out the edge cases that break ordinary notions, accident, illness, moral luck, and sets the bar that any serious law must clear.

1.3 The Hard Cases No One Likes

A theory of fairness that only works on easy days isn't a very useful theory. Philosophers call the brute intrusion of luck into life outcomes moral luck (Nagel, 1979); LoF offers a strictly this-life hypothesis for answering it: even if luck skews events, the lifetime ledger will not remain skewed at closure. The cases below are where good intentions, neat procedures, and clever slogans fail. They aren't edge curiosities; they're places real lives actually go. If a law of fairness is going to mean anything, it must face these without flinching and without changing its terms.

1.3.1 Catastrophe without blame

A toddler with a rare cancer. A teenager paralyzed by a stray bullet. A parent lost to an aneurysm on a sunny morning. There's no villain to prosecute; no procedure can undo what has already been felt. Better distribution and opportunities help before and after, but they don't reach what has already happened inside the stream. Any serious standard of fairness must speak to ledger entries no policy can erase.

Constraint for our law: If fairness is literal, it cannot depend on finding a culprit or making retroactive edits to experience. Fairness must be achieved within a life, in real time, by that stream itself, not by rewinding or by external judgment.

1.3.2 Chronic conditions that do not rebound

Adaptation is real, and limited. Some people live inside pain, treatment-resistant depression, psychosis, or neurodegeneration that never fully loosens its grip. "Most people bounce back" is true, and irrelevant to those who don't. On a ledger, even a small constant pain can out-sum a lifetime of small pleasures.

Constraint for our law: A true law must handle persistent negative drift without assuming spontaneous recovery. If balance is guaranteed, there must be channels of compensation that do not require the world to become perfect first.

1.3.3 Injustice that is never seen or admitted

A quiet assault. A private humiliation. A theft that is legally invisible but psychologically loud. Restorative processes are powerful when they exist; often they don't. What then? Telling someone "one day you'll find meaning in this" is not fairness; it's pressure.

Constraint for our law: Compensation cannot depend on public confession or formal ritual. It must be able to occur without the perpetrator's help, via other lawful channels (dreams, new relationships, art, personal growth, ordinary luck) that register in the ledger for what they truly are: felt counterweights.

1.3.4 Villainy that goes unpunished

A tyrant dies warm and wealthy. A con artist slips every charge. If our theory needs last-minute cinematic justice, it will fail most of history. Aggregate “greater good” calculations can excuse this (“overall happiness increased”); desert-based morals can rage at it. Neither tells us whether that individual’s stream ends with surplus, or whether their victims’ streams end with debt.

Constraint for our law: The target is not moral desert but ledger neutrality. If the law is true, a life that imposed great suffering cannot close with unbalanced surplus pleasure, not because a judge decrees it, but because trajectories that cannot be compensated by closure are never admissible in the first place.

1.3.5 Early endings and tiny ledgers

Miscarriages, stillbirths, neonatal loss: brief streams with few entries. “Balance at the end” risks sounding obscene here. Yet a law that excludes the smallest lives would abandon the most fragile among us.

Constraint for our law: The unit is the unified stream, however short. The concept of balance must be defined with extreme care for tiny T, including wide uncertainty bounds and perhaps the possibility that very short streams are trivially neutral (few or no integrated entries). The law cannot demand compensations that simply don’t fit into the time available.

1.3.6 Identity fog: coma, anesthesia, dementia

When is the “stream” present? Sleep and anesthesia pause it; delirium and dementia fragment it. Ledger math that ignores identity could double-count or erase experiences. But ledger math that treats every flicker of consciousness as a whole new life invites nonsense.

Constraint for our law: Identity rules must be fixed ahead of time (see Chapter 9). When unity is present, the ledger accrues; when it’s absent, the ledger is paused; when streams split or fuse, ledgers split or fuse accordingly under fixed rules. No retroactive combining or splitting after outcomes are known.

1.3.7 Colliding rights, competing needs

Two patients need the only ventilator. A rescue boat can reach one village before the flood takes the rest. Procedural and distributive ethics matter here; they determine how we decide in these terrible situations. But however fair the decision is, one stream will carry more purple beads (suffering) than green.

Constraint for our law: The law must be orthogonal to allocation choices. It cannot make scarcity or tragedy go away. It can only say this: regardless of how resources are allocated, no stream's lifetime ledger is permitted to close in permanent deficit. In other words, even the one who doesn't get the ventilator would somehow have to be compensated by life's end, if the law holds.

1.3.8 When help hurts

A treatment saves a life but leaves chronic pain. A just verdict retraumatizes a victim through public scrutiny. A well-meaning message pressures someone to "find the silver lining." Remedies are necessary; they can also add entries to the wrong side of the ledger.

Constraint for our law: Channels of compensation must be plural and humane. The theory must never prescribe suffering as the medicine. If balance exists, it is more often realized through rest, care, safety, creativity, and connection, not through new wounds.

1.3.9 Survivors, guilt, and the arithmetic of grief

One sibling survives; another does not. One friend escapes the burning building; another perishes. Guilt drops purple beads into the survivor's ledger, sometimes for years. Rituals and therapy can help; fate rarely writes clean endings.

Constraint for our law: Balance cannot require forgetting. It must allow counterweights that coexist with memory; peace is not amnesia. If a ledger reaches zero, it may do so by densifying the good rather than erasing the loss.

1.3.10 Culture, creed, and the stories we tell

Some cultures frame suffering as test, karma, or purification; others frame it as error to be fixed. Stories shape how compensation can come, what counts as repair, what counts as insult. A single theory must not erase this variety.

Constraint for our law: Measurement of experience must be invariant across cultures (Chapter 8), and channels of compensation must be plural enough to fit different life narratives. The law is about ledgers, not scripts; it doesn't prescribe how balance is achieved, only that some genuine balance must occur.

1.3.11 The "nothing happened" day

Many days are neither tragedies nor triumphs; they are errands, emails, light aches, small jokes. Over decades, these "flat" days dominate the integral of life. If a fairness theory only works in crises, it's theater.

Constraint for our law: Our measure must capture low-amplitude, long-duration affect accurately. Tiny drifts, integrated over years, are decisive. Precision in the ordinary is how we avoid needing “magic” in the extraordinary.

1.3.12 The tone we keep.

Writing about fairness at this depth risks two opposite errors: softening pain into “meaning,” or hardening meaning into mere arithmetic. We will do neither. The claim we test is structural, about what kinds of conscious histories are admissible in a world like ours. It never excuses harm and never demands it. Where the world offers room for compensation, people, through compassion, craft, medicine, and ordinary kindness, are often the hands that make it real.

These are the cases a true law must bear. In the next section, we name a widespread temptation, “life tends toward fairness,” and explain why a tendency is not enough. A tendency can leave some lives behind. A law, if it exists, cannot.

1.3.13 Where we go next:

Having faced the cases that break easy theories, catastrophe without blame, chronic non-rebound, unseen harms, unpunished villainy, identity fog, and more, we now draw a bright line between a tendency and a law. In 1.4 we will show why averages, drift back to baseline, or opponent-process “after-effects” cannot guarantee a neutral ending for each stream, and why only a true constraint could do that work.

1.4 Why “Tendency Toward Fairness” Isn’t Enough

It is tempting to say, “Life tends to balance out.” Psychology describes adaptation; physiology describes opponent processes; statistics promises regression to the mean. These are real and important. But a tendency is not a guarantee, and our target, life-level fairness for each conscious stream, would require a guarantee. This section explains why a mere tendency falls short.

1.4.1 Averages do not rescue individuals

Many findings are population-level: most people drift back toward baseline mood after gains and losses; most recover some function after injury; most settle into ordinary days eventually. This is the well-known “hedonic treadmill” in action: our happiness often slides back toward a personal set-point even after extreme ups or downs. A classic study comparing lottery winners and accident victims is often cited in this context (Brickman, 1978). But even when the average looks stable, individual outcomes vary widely. “Most” is not “each.” A law of fairness must speak to this stream, not just to an average of streams. If ten people rebound and one does not, adaptation can be true and the law could still fail for the one. Rule we keep: Evidence about means is not evidence about closure for every ledger.

1.4.2 Time slices are not lifetimes

Adaptation is usually demonstrated over weeks or months, not from birth to the Death of Mind. A person can return to baseline after a breakup and never return after a head injury. Tendency models describe local drifts. The Law of Fairness makes a claim about the integral over an entire life. The first cannot substitute for the second. Rule we keep: What drifts back in a season may still sum to imbalance in a life.

1.4.3 Regression to the mean is not compensation

Statistical regression explains why extreme scores tend to be followed by less extreme scores when there’s noise. It says nothing about felt compensation. If you shoot two arrows and the first falls long and the second is closer, statistics smiles, yet the target wasn’t hit. Calling noise “healing” confuses math with meaning. Rule we keep: Falling back toward average is not the same as adding counterweights to a ledger.

1.4.4 Opponent processes are local, not terminal

Opponent-process theory says a surge in one direction triggers a counter-surge (fear to relief; drug high to crash). Notably, pain-offset produces measurable pleasure: the abrupt cessation of pain engages reward circuitry and yields relief greater than a mere return to baseline (Leknes, 2008), a clear local counterweight without implying

guaranteed lifetime neutrality. But those counters can be small, brief, or asymmetric. They can also wear out (tolerance) or fail entirely (chronic pain, major depression). Taquet (2020) provides empirical support: they find that individuals with low mood or a history of depression show impaired mood homeostasis, meaning their mood is less effectively stabilized. Rule we keep: Local counter-reactions do not imply final neutrality.

1.4.5 Survivorship bias hides the tails

We meet many people who “came through it.” We meet fewer who did not, because they dropped out of the sample (out of the workforce, out of social circulation, or out of life). Stories of resilience are true and overrepresented. A theory that leans on visible rebounds risks survivorship bias, missing the ledgers that stayed dark. Rule we keep: Resilience data must be corrected for those who are missing (the invisible failures to rebound).

1.4.6 Optimization isn’t fairness

Brains and cultures optimize: they reduce prediction error, conserve energy, maximize reward under constraints. In other words, the mind functions as a prediction-error minimizer (Friston, 2010) and energy regulator (Sterling, 2012), aiming for stability rather than equity. Optimization can make life work; it doesn’t ensure life is fair. A thermostat that holds temperature steady doesn’t compensate yesterday’s cold; it just keeps today comfortable. Rule we keep: Stability or efficiency says nothing about balanced endings.

1.4.7 “On average, neutral” still allows tragedies and windfalls

Suppose a whole population’s mean ledger is near zero. That’s compatible with wide variance: some lives end far negative, others far positive. If fairness were merely a population property, that might seem okay. But if fairness is a life-level property, as it must be if we take the child, the rockstar, and the monk seriously, then variance matters more than the mean. Rule we keep: A fair society is not the same as fair lives.

1.4.8 Hidden circularities

Some claims that “things balance out” smuggle the conclusion into the premise. For example: only people who report balance are sampled, those who don’t are labeled “outliers” and excluded; or a clinic’s follow-up finds improvement because only those who improved kept coming back. If we choose our samples, definitions, or endpoints after seeing outcomes, we can manufacture a false “tendency.” Rule we keep: No post hoc endpoints. We won’t redefine success once we’ve seen the data.

1.4.9 What a law must add

A law forbids certain endings; it does not merely discourage them. In this book's proposal, trajectories that would make terminal neutrality impossible are progressively rendered inadmissible as a life unfolds. We hypothesize that this pruning operates through the Queue System (QS), not as an external director or moral accountant, but as a constraint on the evolving field of thoughts, choices, and actions available to a unified conscious stream. As ledger imbalance grows or as the remaining horizon narrows, the set of realistically selectable paths shifts. Options that would lock in permanent surplus or deficit become harder to sustain, while compensatory paths become more salient or reachable.

A tendency lets unfair endings happen; a law disallows them. This kind of strict closure constraint is unusual in biology. Most systems maintain homeostasis, stability within bounds, rather than enforcing a lifetime-integrated neutrality condition. LoF posits something more exacting: a higher-order feedback structure in which accumulated imbalance increasingly shapes what is thinkable, doable, and sustainable, so that no extreme surplus of pleasure or pain can remain structurally uncorrected by the time the Death of Mind is reached. A tendency might allow a few lives to end wildly imbalanced; a law would not permit such endings within its admissible histories. The difference is everything.

Plain version: A tendency says “usually.” A law says “never, under lawful conditions.”

1.4.11 The ethical reason “tendency” fails

“Tendency toward fairness” sounds kind. In practice it invites resignation: most people bounce back, so we needn’t try so hard. The law we will test forbids that conclusion. If balance is real, it happens through care, repair, art, rest, and justice, not by magic. If balance isn’t real, those are still our duties. Either way, “tendency” is the wrong word for what matters most.

1.4.12 Where we go next:

In 1.5 we explain how to read this book, two voices (plain and technical), visual conventions, fail-pattern boxes, and ethics first, so you can follow the argument without losing the human thread. Then we turn to Chapter 2, where we argue why feelings, not things, are the final currency any fairness claim must settle.

1.5 How This Book Works

This book speaks in two voices at once, one for any thoughtful reader, and one for researchers who want the machinery. The claim never changes between voices; only the level of detail does. Here's how to navigate so the thread stays clear.

1.5.1 Two tracks, one claim

Main text: Plain language narrative, stories, intuitive examples, comprising the main chapters. You can read only this and grasp the whole argument.

Research Notes: Gray sidebars containing definitions, equations, study designs, and references. They show how each idea could be tested and what would count against it. They're skippable without losing the story, and sufficient on their own if you want the technical backbone.

1.5.2 Evidence standards and Fail patterns

Every empirical claim in this book is paired with Fail patterns – data that would weaken or refute the law. (Look for red-bordered “Fail pattern” boxes.) The canonical signatures we test throughout are:

- Dream counterweights after difficult days
- Horizon scaling as end-of-life nears
- Residual imbalance signals in affect-regulation networks, after accounting for other factors
- Variance compression of life ledgers as streams approach closure

Absence of these patterns, under adequate methods and statistical power, counts against the Law of Fairness. Most importantly, if any unified conscious stream reaches the Death of Mind with a clear residual imbalance beyond the defined neutrality band, the Law is false.

1.5.3 Ethics first, everywhere

- End-of-life observations will never interfere with a person’s comfort or a clinician’s judgment.
- Simulation work uses only non-sentient agents (no conscious creatures harmed or trapped).
- Personal data are handled with dignity; examples are anonymized or illustrative unless explicitly noted.

- Balance at the end is never an excuse for suffering now. All normal duties of care and compassion remain, no matter what any theory suggests.

1.5.4 How to move through the book

- Human arc: Read Parts I, III, VI, IX, and XI for the narrative, human-impact perspective (skipping deeper technical detail in between).
- Technical arc: Include Parts II, IV, V, and X (the core theory, measurement, edge cases, and experimental methods), plus Appendices A–F for full technical details.
- Time-limited: If you only have time for the highlights, read Chapters 3 (the Law itself), 7–8 (measuring feeling), 10–12 (evidence tests), and 23 (hard objections and answers).

1.5.5 What changes if you disagree

Different skeptics can start at different places. This book offers guidance:

- *Think feelings can't be measured? Check out Chapters 7–8 (on measurement) and Objection 23.1 (on the challenge of quantifying experience).*
- *Believe "most people adapt, that's enough"? See Chapter 14 (hedonic adaptation vs. a strict law) and revisit Section 1.4 above.*
- *Think the brain is just a prediction-error minimizer? See Chapter 15 (predictive coding's take on these phenomena).*
- *Worried we're sneaking morality into physics? See Chapter 4 (constraint, not purpose) and Chapter 7's discussion of keeping the science honest.*

1.5.6 Where we go next:

Chapter 2 establishes why feelings, not money, status, or even “objective” outcomes, are the final currency any fairness claim must settle. From there, we formally state the Law, build the measuring stick, name the tests, and invite the evidence to prove us right or prove us wrong.

Chapter 2 — Feelings as the Final Currency

Imagine two patients facing a hard choice. Alice can undergo an aggressive treatment that might extend her life by a year, but it will leave her in constant pain. Binh, in contrast, opts for comfort-focused care with no heroic interventions; he may live somewhat less long, but his remaining time will be relatively peaceful. By the numbers, Alice “gains” more days than Binh. Yet who receives the fairer outcome? If fairness were measured by simple metrics like years or procedures, Alice’s outcome appears better. But when measured by lived experience, the comfort or torment each person actually feels, the picture reverses. We intuitively recognize that an extra year of suffering may be worth less than a shorter time spent in relative ease. Fairness, then, cannot be judged by external inputs alone; it ultimately turns on the quality of the experiences those inputs produce.

Feelings over figures. We encounter this truth in everyday life. Two colleagues can earn the same paycheck, yet if one lies awake each night anxious and sleepless while the other rests soundly, their lived outcomes are not equal. Money, status, awards, even years of life are means to an end; they are not what living ultimately delivers. What life delivers is felt experience, the comfort and ache, fear and relief, boredom and interest, humiliation and dignity, peace and wonder that fill our days and nights. A theory of fairness that never reaches this level can declare victory on a spreadsheet while the people inside the numbers lose. In this chapter, we take a simple but demanding step: we treat felt experience as the final currency that any life-level notion of fairness must settle.

This does not mean resources, rights, and opportunities are irrelevant. It means their importance derives from how they translate into experience. A perfect distribution of goods or chances means little if it leaves one person quietly miserable and another content. Fairness, if it is to mean anything deep, must answer a single question: how did it feel to live that life? Only by targeting feelings can we speak about fairness in a way that is both humane and testable. Under the Law of Fairness, the system must balance each life’s felt ledger by the end, not merely tend to do so on average. That is a bold claim, and to state it rigorously we must clarify what we mean by feelings and how they can be counted.

We approach these tasks in six steps. First, we explain why feelings, not just things, are the right target for a fairness calculus. Second, we ask whether subjective feelings can be measured without trivializing them and outline a careful approach. Third, we examine how feelings guide decisions through the somatic marker hypothesis, showing that an experience-based perspective appears in real choices. Fourth, we sketch a concise tour of the biology of feeling, anchoring the argument in neuroscience without overwhelming

detail. Fifth, we set non-negotiable requirements for any measure of feeling: it must function across cultures in meaningfully comparable ways and it must never violate human dignity or consent. We name invariance standards and ethical red lines, along with fail conditions that would force us to abandon the project. Finally, we preview the life ledger concept that will formally track the running balance of experience in later chapters, introducing Hedonic Composite Units (HCU) as the currency for tallying pleasure and pain over time.

Throughout this chapter we use two voices. The main text remains in plain language and intuitive examples. Short Research Notes provide technical scaffolding, how we construct a Hedonic Composite Index (HCI) to estimate momentary felt experience, how we test measurement invariance, and how uncertainty is handled. You can ignore those notes and still follow the argument, or focus on them to see how the mathematics respects the humanity of the problem. The claim does not change between voices; only the level of detail does.

What you'll get from this Chapter:

- Why feelings are the true yardstick: You'll see why fairness must ultimately be evaluated in terms of happiness and suffering, not wealth, luck, or other external scores. A life can have every outward advantage and still be deeply unfair on the inside, and the reverse is also true.
- Measuring the “immeasurable,” carefully: You'll learn how subjective feelings can be systematically estimated without trivializing them. We introduce a cautious composite metric (the HCI) and explain how self-reports, behavior, and biology can converge into a single estimate with explicit uncertainty.
- Biological footing of feeling: We provide a concise tour of how brain and body generate feelings. This shows that feelings are not mystical; they have a physical basis we can track, even though no single signal captures a whole life.
- Ethical guardrails in research: You'll see the standards we set before measuring anyone's experience. Cross-cultural comparisons require strict measurement invariance tests. If those fail, we restrict conclusions accordingly. Relief is a systems variable; comfort and dignity override data collection.
- The lifetime ledger concept: You'll preview how we sum a life's experiences into a ledger. We introduce the idea of integrating moment-by-moment affect (HCI) over time to yield total HCU, and explain what it would mean for that ledger to close neutral at the end of consciousness.

Subsections in this Chapter:

- **2.1 Why Feelings, Not Just Things** - Inputs are means; experience is the end. Two identical paychecks, two different nights of sleep, proof that the same external gain can yield unequal lives. This subsection centers feelings as the only currency a life actually spends and shows why meaning itself appears as feeling.
- **2.2 Can Feelings Be Counted?** - No single meter can read a life, but a cautious composite can. We introduce a multi-channel approach, self-report, behavior, physiology, brain signals, and dreams, calibrated within person and linked across people, with uncertainty carried into every conclusion.
- **2.3 Somatic Markers** - Before a decision becomes a sentence, it is a tug in the body. We show how learned bodily signals bias choices under uncertainty and why those markers matter for any mechanism that could keep a life's ledger compensable without teleology.
- **2.4 Affective Neuroscience in One Page** - Feelings live in a body-brain loop. We outline key circuits and neuromodulators that make experience lawful enough to measure and connect them to signatures tested later.
- **2.5 What We Will Never Do** - Dignity over data, always. We set non-negotiable red lines: no harm induction to “prove balance,” end-of-life comfort before measurement, privacy by design, and transparent preregistration.

Where we go next:

We begin at the only place fairness can cash out. Section 2.1 argues, in clear terms, why feelings, not money, status, or even years, are the final yardstick for a fair life. From there, we build the careful meter and the ledger that make the law auditable.

2.1 Why Feelings, Not Just Things

If fairness is to say something true about how lives actually go, it must address what lives deliver. Life does not deliver money in an account, a title on a door, or an extra year on a calendar. Those are means. What life finally delivers is what it was like to be the person who had those means, moments of comfort or ache, fear or relief, humiliation or dignity, boredom or interest, peace or wonder. Feelings are not decorations on outcomes; they are outcomes as lived.

2.1.1 Inputs are means, experience the end

Two people receive the same raise. One sleeps better, laughs more, and feels safer walking home. The other's anxiety spikes with new obligations. The same input yields different lives. Two patients gain six months from a treatment. One spends those months lucid, pain-controlled, with friends and music; the other spends them nauseated, frightened, and isolated. The same added time does not mean the same thing. When we treat things as ends in themselves, money, status, years, we mistake the map for the terrain. The terrain is lived experience.

2.1.2 Preference satisfaction isn't enough

It is tempting to say fairness means people get what they prefer. But preferences often mispredict how states feel once we inhabit them. We mis-forecast our affect, cling to options that later feel hollow, and avoid choices that later feel deeply right. People can prefer the familiar even when it produces a worse felt life. The Law of Fairness cannot be framed in terms of wishes fulfilled; it must be framed in terms of what actually occurred in experience. Research on hedonic adaptation shows that even dramatic life events often have less long-term emotional impact than expected, as people drift back toward a baseline. Our intuitions about what we want can therefore mislead our sense of what will improve lived experience.

2.1.3 Meaning shows up as feeling

Insisting on feelings does not reduce life to mood. Meaning, a sense that one's life hangs together and actions fit values, registers as feeling: relief when a promise is kept, calm when a decision aligns with commitments, warmth when belonging is real. Existential goods are not outside experience; they are among its most important qualities. By focusing on felt experience, we do not exclude meaning or virtue; we include them in the only currency that ultimately matters: what it feels like to exist.

2.1.4 Lawful signals: how to measure

We need not invoke mysticism to take experience seriously. Feelings have a biological address in interoceptive maps, control hubs, valuation circuits, neuromodulator systems, and autonomic pathways. Sleep and social contact regulate them. This lawfulness makes cautious measurement possible. We will never obtain perfect meters, but we can construct convergent ones that are reliable enough to test a law.

2.1.5 Ethics sharpened by the target

If experience is the end, our duties sharpen. We do not congratulate ourselves for procedures that are fair on paper if they leave people worse off in fact. We do not praise longevity purchased at the cost of torment. We design institutions not merely to allocate inputs, but to improve lived days, to reduce avoidable fear and pain, and to make room for safety, dignity, interest, and rest.

2.1.6 Why this matters for a law

A law that claims life-level fairness must evaluate life-level outcomes in the currency that counts. That currency is felt experience integrated over time. Only with that target can we state a claim that might be true or false for every Unified Conscious Stream rather than for population averages or inputs. Only with that target can we ask whether each life ledger closes neutral at the Death of Mind. Section 1.5 introduced our claim in one sentence; focusing on feelings makes that claim concrete. Science suggests at least a tendency toward balance in affect: nervous systems seek homeostasis, and known mechanisms counteract extremes. Intense emotions often trigger opponent-process effects, and people frequently return toward a personal baseline after major events. Negative emotions tend to fade faster than positive ones, shaping memory over time. These patterns do not guarantee a neutral lifetime sum, but they hint that biology resists permanent extremes. The Law of Fairness extends that idea to its strongest form and treats it as a testable hypothesis. We will continually ask what evidence would confirm or refute such balance.

2.1.7 “Subjective” doesn’t mean arbitrary

Objection: “Isn’t this subjective?” Yes, and that is precisely the point. The subject is the one who suffers and enjoys. Subjectivity does not mean arbitrariness. We use composite, invariant measures, self-report where possible, behavior, physiology, neural signals, and dream affect, to triangulate a latent estimate with explicit uncertainty. The number is not the life; it is a disciplined way of respecting each life in its own terms. If measurements cease to mean the same thing across people or contexts, we do not paper over that failure.

2.1.8 A quick decision test

You can apply the final-currency rule to any decision: will this change improve what it is like to be the people who live under it? If the answer is unclear, count more than inputs. Ask about sleep, safety, companionship, pain, interest, shame, and peace. Those answers are what a lifetime ledger records.

2.1.9 Where we go next:

Having fixed the target, we turn to the meter. In 2.2 we outline how feelings can be counted cautiously, composites over single gauges, within-person calibration before between-person comparison, and equivalence bounds that keep us honest.

2.2 Can Feelings Be Counted?

Short answer: yes, imperfectly, cautiously, and well enough to test a law. “Counting” feelings does not mean reducing a life to a single digit. It means using many imperfect clues, together, to form a humble estimate of how moments felt, and carrying explicit uncertainty all the way to any conclusion. This section explains how.

2.2.1 The composite approach

No single meter “reads” experience. Self-reports are honest but noisy. Physiology is objective but ambiguous. Brain signals are lawful but indirect. Behavior is observable but context-bound. Rather than choosing one channel, we combine them. We call the result the Hedonic Composite Index (HCI): a latent estimate of momentary net affect built from converging indicators. (By net affect we mean the overall emotional state at a given moment, balancing positive and negative feeling.) The latent HCI is not any one measure; it is whatever hidden value best explains all of the signals at once. Crucially, the HCI always travels with error bars (uncertainty), so we never overclaim precision.

Typical ingredients (sampled over time) for HCI include:

- Self-report: Brief in-the-moment ratings; plus private, incentive-compatible tasks to reduce any impulse to “fake good.”
- Behavior: Choice under risk, persistence vs. avoidance, micro-latencies in response, effort expended on tasks.
- Physiology: Heart-rate variability, electrodermal activity (sweat response), respiration patterns, pupil size, facial EMG (minute muscle movements).
- Neural: EEG rhythms and event-related potentials; occasional fMRI snapshots for ground-truth regions (like insula, ACC, vmPFC) in smaller sub-studies.
- Dream affect: Morning dream reports coded for valence and intensity (see Chapter 10 for why we care about dreams).
- Contextual priors: Pain diary entries, medications, sleep quality, social contact, even weather – to help explain fluctuations.

Each channel is weak alone. Together, with the right math, they form a reliable compass.

2.2.2 Within-person first, between-person second

Feelings are idiographic: your “+3” may not be my “+3.” We therefore calibrate the HCI within each person first, then link people onto a common scale. Within-person calibration means each person completes a short anchor battery that includes a few standardized experiences with known affective impact (for example: a brief cold pressor

test, a CO₂ breath-hold, a social inclusion vs. exclusion vignette, a music-induced chill, a small monetary reward task). From these, we learn that person's response range and emotional "gain" (how strongly their signals respond). Next, for between-person linking, we use anchoring vignettes and Item Response Theory (IRT) methods to align scales across individuals, and we explicitly test for measurement invariance; in plain terms, we ask "Do our items and signals behave similarly across different groups of people?" We require at least metric invariance (comparable factor loadings/slopes) before comparing magnitudes between people, and we report where scalar invariance (comparable intercepts/baselines) holds or fails. If full invariance fails, we restrict claims to within-person changes or incorporate adjustment factors for group differences, propagating the added uncertainty. The result is not a mythical perfect scale; it is a careful bridge between different nervous systems.

2.2.3 Sampling the stream without drowning

Experience flows continuously, but we can't measure every second. To track it responsibly, we sample little and often:

- Ecological Momentary Assessment (EMA): 3–6 ultra-brief check-ins per day via phone or watch, timed to avoid bias (random within preset windows, and silent if the person is moving or under stress to avoid interference). Each EMA might ask "How do you feel right now?" on a simple slider, plus one or two quick taps for context.
- Passive signals: Wearables capture background physiology (HRV, EDA, activity) and phone sensors provide context proxies (light exposure, movement, social app use as a proxy for interaction) to fill some gaps between active reports.
- Sleep and dreams: A short morning questionnaire asks for a two-minute sketch of dream affect and mood after waking. In some sub-studies, an optional home EEG headband logs sleep stages (REM/NREM) to correlate with affect processing overnight.

We do not need every second of data. We need representative glimpses and then principled interpolation. In practice, we use state-space models that respect the known features of affect dynamics (the relative slowness of mood drifting vs. the spikiness of acute events) to estimate what happened in between samples.

2.2.4 Turning many clues into one estimate

Under the hood, we fit a Bayesian state-space model where the hidden state is momentary net affect, denoted F(t). The observed channels are noisy functions of F(t) (each with its own quirks and error distributions). The model learns, for each person, how

strongly each channel tracks the hidden state. It then outputs an HCl time series, essentially our best estimate of $F(t)$ over time, with uncertainty bands at each moment. (Full details of the model, priors, and computation appear in Chapters 7–8 Research Notes.) (In effect, this approach models experience as a continuously changing internal state, a state-change formalism that focuses on moment-to-moment dynamics rather than any static labels.)

Plain speech: Imagine several shaky compasses in your pocket. Each wobbles for different reasons (one is affected by temperature, another by vibration, etc.). If you learn the quirks of each compass, you can combine them in a smarter way than just taking a simple average, weighting those that are steadier, accounting for known biases, and you will also know how confident to be in the combined reading at any given time. HCl is that combined reading of the emotional “needle.”

2.2.5 From HCl to units you can add up

To integrate over a life, we need a unit of measurement. We define Hedonic Composite Units (HCU) as the integral of HCl over time, effectively, “emotion-time” on an anchored scale. In practical terms, we first anchor the HCl scale so that a one-point change has concrete meaning. For example, a one-minute increase of +1.0 on the HCl might be defined to equal the affect boost of a median music chill or a mild social inclusion experience. Similarly, -1.0 for one minute might correspond to a minor social rejection or a physical discomfort of short duration. We use multiple such anchors (pleasant and unpleasant) so no single task defines the whole ruler. Once HCl is calibrated in these meaningful units, we can sum it over time. One minute at +1.0 HCl gives +1 HCU; ten minutes at +2.0 HCl gives +20 HCU, and so on. Critically, this time-integration avoids the biases of memory. Humans tend to neglect duration and judge an experience by its peak and ending moments rather than its total sum (Fredrickson & Kahneman, 1993). By summing every moment of HCl into HCU, we ensure that a life’s ledger isn’t misled by rosy retrospection or painful endpoints; it reflects the true cumulative experience. Likewise, negative HCl values subtract HCU. For example: if your HCl was +2 for 10 minutes, that adds +20 HCU; later, an HCl of -1 for 30 minutes subtracts -30 HCU; the net for that period would be -10 HCU (with uncertainty carried from each measurement). Because HCl is person-calibrated first and then linked across people, +10 HCU is meant to reflect roughly the same magnitude of felt change for you as for someone else (within the stated error).

2.2.6 Guarding against the usual traps

Any attempt to quantify experience must be on guard for well-known pitfalls. We design our protocol to preempt or catch these:

- Response bias: We intermix private implicit tasks and physiological/behavioral channels to counteract any self-report distortions. Participants can't game the composite easily, because it's not just "fill out a survey."
- Cultural or linguistic drift: All self-report items are developed by forward-back translation in multiple languages and checked for *differential item functioning*. We use anchoring vignettes across cultures to align interpretations of the scale.
- Device/hardware bias: Wearables are cross-calibrated. If we upgrade devices, we run them side-by-side for an overlap period to recalibrate the new data to the old scale.
- Missing data: We expect gaps. The modeling framework treats missingness explicitly (filling in via the state-space model when possible), and we audit any systematic patterns in dropout or missing data to ensure they aren't themselves the signal.
- Reactivity: Could measuring people's feelings change those feelings? We minimize this risk with a low sampling frequency, vary prompt times so it's not predictable, and include occasional "silent weeks" to check whether the act of measuring is altering the experience.
- P-hacking: All primary outcomes, equivalence bounds, and analysis plans are preregistered. We lock these decisions before looking at data. Any deviations are logged publicly with a justification. No fishing expeditions with hindsight.

2.2.7 Equivalence bounds and honest uncertainty

A key commitment of our approach is that we never declare "Neutral!" just because a mean ends up near zero. We define equivalence bounds, a narrow band around zero (in HCU units), before seeing outcomes. At the end of the study, if the lifetime ledger's entire confidence interval lies within that band, we count it as neutral (balanced); if the interval lies entirely above the band or entirely below it, non-neutral; if it straddles the bound, inconclusive. This forces us to be honest. It keeps the law falsifiable and our language disciplined. We are effectively saying, "Only if we are confident the ledger is very close to zero (within a tiny margin we agreed on ahead of time) will we call it balanced. Otherwise, we either have evidence it's off or we don't know." In technical terms, this is the principle of equivalence testing rather than standard null-hypothesis testing, treating a near-zero result as meaningful only if it falls inside a pre-specified small interval.

2.2.8 Ethics: measure with dignity

All this measurement must occur in a way that honors the person. We minimize burden (the app is lightweight and optional if someone feels overwhelmed), allow opt-out windows at any time, and prioritize comfort in clinical settings. Raw personal data are encrypted at rest; only de-identified, aggregated results leave the device (with differential privacy techniques adding a little statistical noise to protect individuals). End-of-life telemetry, when we do it, is non-interfering (it never dictates care) and is staffed by clinicians first, researchers second. And for any simulations we run to test ideas, we include a non-sentience safeguard; we simulate only systems that, under our best current understanding, are not conscious and therefore not capable of suffering (more on this in Section 2.5.5).

2.2.9 When counting would fail (and we would stop)

We would reconsider the entire program if we found that any of the following fail patterns persist despite our best methods:

- Composite not invariant: Composite measures cannot achieve even partial invariance across cultures after serious design effort (meaning we can't trust we're measuring the same construct everywhere).
- No signal in body or brain: Neural or physiological channels have no stable correlation with careful experience sampling; in other words, if brain and body signals just don't track self-reported experience in any reliable way.
- Calibration drift: Within-person calibration drifts so wildly over time that no meaningful cumulative ledger can be constructed (today's "+2" ends up meaning nothing like yesterday's "+2" for the same person).
- Irreducible uncertainty: The uncertainty around $L(T)$ remains so wide for most participants that our equivalence testing becomes vacuous (a confidence interval so large it never falls fully inside or outside the neutral zone).

If these occurred consistently, they would indicate that our meter isn't working. A law is only as good as the meter that can test it. In such a case, we would publish the failure, halt the project, or rethink it fundamentally. We do not "prove the law" by ignoring that our measurements fell apart; we would take it as evidence that the Law of Fairness (in its current form) might be untestable or false.

2.2.10 Why "counting" is the humane choice

Some worry that reducing life to numbers is cold or mechanistic. But avoiding clear numbers does not protect dignity; it protects vagueness. Clear, cautious numbers,

paired always with uncertainty and rigorous ethics, let us test whether the most consequential moral claim we can make about a life is true: that its felt balance closes neutral. If the claim stands, the numbers will help us see where and how balance is realized. If it falls, the numbers will help us learn why and what to do instead. In either case, quantification (done humbly) is a tool for understanding and improving lived experience, not an end in itself.

2.2.11 Where we go next:

We have a cautious composite and a principled way to turn moments into units. Next, in 2.3, we descend from meter to mechanism: how somatic markers, the body's learned "yes/no," bias choices under uncertainty and could make compensable paths feel reachable without any cosmic steering.

2.3 Somatic Markers

Before a decision appears as a sentence in your head (“I think I’ll do X...”), it often arrives as a tug in your body, a quickening chest, a hollow in the stomach, a warm yes, a tight no. Neurologist Antonio Damasio called these bodily signals somatic markers: learned tags that couple situations with visceral states and, in doing so, tilt our choices. They are like post-it notes from your past experiences, attached to options: “This situation feels bad; avoid it” or “This one feels promising.” Somatic markers are shortcuts built from experience. They do not guarantee wisdom (we can have misleading gut feelings), but they make some futures more reachable than others by making those futures feel a certain way.

2.3.1 The idea in one sentence

Somatic markers are bodily feelings linked to options that bias decision-making toward or away from those options, especially under uncertainty and time pressure. They are the emotional valence (positive or negative value) of our options, experienced as intuition.

2.3.2 Why this matters for fairness

If fairness must be realized inside lives, not just by decree on paper, then how choices are guided matters. Somatic markers are a primary bridge between the ledger of past experience and the next action you take. They make the past felt in the present. This is exactly the kind of mechanism by which balance could be pursued without any mystical teleology: instead of the universe steering us, our own accumulated feelings nudge us toward trajectories that might even out our experience. In this book’s language, somatic markers are one way the Queue System (QS) appears from the inside. You do not consciously see the QS pruning your options; you just feel some options as a tugging yes or no among the ones that remain. A strong negative marker can make a harmful choice feel aversive before your reasoning fully catches up. A positive marker can make a reparative, healthy choice feel available, where a purely calculative (but affect-blind) brain might stall or not see the opportunity. In short, markers inject the wisdom of experience (and the constraints of the QS) into day-to-day decision loops.

2.3.3 Classic observations (lab and life)

Two oft-cited examples illustrate somatic markers:

- Iowa Gambling Task: In this psychology experiment, people gradually learn to prefer “safe” decks of cards over “risky” decks, but crucially, their bodies learn before their brains do. Measurements like skin conductance start responding to bad decks even when participants can’t yet verbalize which decks are bad.

Patients with ventromedial prefrontal damage (who lack these gut responses) often fail to learn the task — they keep choosing from bad decks despite mounting losses.

- Everyday choices: The text message you draft and then delete, the dark shortcut home you decide not to take, the invitation you accept at the last minute — many such everyday decisions are influenced by a somatic “yes/no” that kicks in faster than explicit reasoning. You might say “I have a bad feeling about this,” without immediately being able to state why.

These patterns support a simple point: bodily feeling can carry useful information about long-run outcomes even when our explicit reasoning hasn’t caught up. Our ledgers of experience inform us, through somatic markers, in ways we don’t always recognize consciously.

2.3.4 Where in the body–brain loop?

Several systems participate in somatic markers:

- Interoception (Insula): The insular cortex maps the state of the body (heartbeat, gut tension, temperature, etc.) and helps generate the felt sense that “something is happening in me.” If this mapping is disrupted (through lesions or numbing), feelings flatten, a clue that insula is central to how somatic signals become conscious.
- Valuation (vmPFC/OFC): Ventromedial/orbitofrontal prefrontal areas link contexts to value. They integrate somatic inputs into the decision process. Patients with damage here often know the facts but can’t feel the stakes; they struggle with decisions because the options don’t feel different.
- Control/Conflict (ACC and rIFG): The anterior cingulate cortex (ACC) registers conflict and error likelihood (“this might go wrong”), recruiting effort or inhibition. The right inferior frontal gyrus (rIFG) is a “brake” that stops habitual actions. When a somatic marker says “Not this,” ACC and rIFG help slam the brakes or muster the will to avoid that choice.
- Autonomic partners (Vagus and Sympathetic): These are the body’s broadcasting channels for markers. A vagus-mediated slow-down or a sympathetic jolt (racing heart, sweating) is part of the marker signal. High heart-rate variability (strong vagal tone) often means more nuanced emotional responses; sympathetic surges accompany strong “uh-oh” or “yes!” signals.

You don't need to memorize this anatomy to grasp the idea; the point is that markers are not metaphors or magic. They live in flesh and blood. They have locations and pathways, which is why we can measure them (next section).

2.3.5 Markers and the lifetime ledger

Somatic markers encode prior entries in the ledger. If certain choices led to a lot of purple beads (unpleasant affect) in your past, similar choices later will tend to carry a purple emotional tint; if certain contexts added green beads (pleasant affect), similar contexts later will feel inviting. Over time, this mechanism makes the organism “credit-sensitive” without any spreadsheets. Green markers (signals of safety, kindness, fit) tend to open trajectories where positive counterweights can emerge without drama, think repair, reconciliation, rest. Purple markers (signals of hazard, shame, depletion) tend to close off trajectories that would add debt faster than compensation could keep up. In this way, if the Law of Fairness holds, somatic markers would be one way the system gently biases life paths toward ones that are compensable.

To be clear, the Law of Fairness does not claim that markers are always correct or good. It claims that lawful streams (ones that do balance out) are those in which, on average, markers, especially near the end of life, bias selection toward compensable futures. In other words, if the law is true, we expect to see that as people’s time horizons shrink, their gut feelings increasingly nudge them toward choices that help balance their ledgers. (We will revisit this in Chapter 6 on the shadow price of remaining time and in Chapter 5 on the Queue System’s decision dynamics.)

2.3.6 How QS might use markers (without mystique)

Think of the Queue System (the LoF mechanism from Chapter 5 onward) as a constraint-weighted filter on your possible actions. It doesn’t add a new magical force; it tweaks the weights on your existing decision processes. Somatic markers are what that weighting feels like to you, the agent:

- As the subjective time horizon H shortens, the “shadow price” of future compensation opportunities λ rises (we’ll formalize this in Chapter 6).
- Options that threaten terminal non-neutrality (i.e. could leave a permanent negative imbalance) acquire stronger purple markers (an instinctive aversion). Options that make compensation likely get stronger green markers (an instinctive attraction).

From the inside, you experience this as “I just can’t bring myself to do that” for certain harmful choices, or “I feel pulled to do this now” for certain reparative choices. From the

outside, a scientist sees ordinary decision circuits doing their usual jobs, but with a QS-residual influence: after modeling standard factors (like the immediate utility of choices and general conflict/arousal levels), there is leftover variance in choice that is explained by how feasible compensation is if that choice is taken.

No angels, no teleology, just felt weights that bias which paths are more likely, in line with the fairness constraint. If the LoF is real, somatic markers would be one medium through which it quietly exerts its influence.

2.3.7 What we can measure now

Markers are measurable enough to matter. We already incorporate them as channels in HCI and as moderators of choices. Concretely, current technology allows:

- **Physiology:** Heart-rate variability (HRV) and electrodermal activity (EDA) track the ebb and flow of marker arousal. Respiration patterns, pupil dilation, and even facial muscle micro-movements (EMG) change with “gut feelings.”
- **Behavior:** Avoidance or approach latency (how quickly someone withdraws from or engages with a stimulus), go/no-go task performance (inhibition ability when feeling a signal), micro-corrections in movement, and eye-tracking patterns (dwelling longer on options that feel safe or glancing away from those that feel wrong).
- **Neural:** Brain responses in insula, ACC, vmPFC to conditioned stimuli can be recorded via EEG/fMRI. Error-related signals (like the error-related negativity in EEG or certain frontal theta rhythms) indicate when a marker might be raising a flag. Frontal midline theta activity, for instance, is a known index of conflict monitoring (ACC) and goes up when you have to override a gut feeling.
- **Self-report:** Quick “gut check” sliders, separate from liking. For example, we ask people to rate how much they lean toward or away from an option, which often captures a visceral sense of comfort or aversion distinct from how much they think they like it logically.

In building HCI (Chapters 7–8), these marker-related signals enter both as part of the moment-to-moment affect estimate and as inputs into models of choice. We thereby acknowledge that how things feel viscerally can alter what people do next.

2.3.8 Ethical use and misuses we will not allow

Because somatic markers influence decisions, they can in principle be manipulated, which is a known tactic in advertising (“gut” branding), social media outrage algorithms, or even coercive persuasion. Our research program draws a strict boundary here:

- No inducing trauma for “balance”: No protocols that intentionally induce harm (trauma, panic, etc.) just to see if a “compensatory” response follows. We are not in the business of creating suffering to test whether it balances (see 2.5.2 for our stance against such harm induction).
- No implanting phobias: No attempts to implant lasting markers through aversive conditioning. We won’t, for example, shock someone in a lab to try to instill a new fear that we then watch them “compensate” for.
- Transparent consent: Full transparency and opt-in consent for any study that even brushes on learned biases. If we explore how markers form or change, participants will know *exactly* what stimuli and goals are involved.

Somatic markers should help us understand fairness, not become a tool for taking advantage of nervous systems. We treat any potential dual-use (e.g. using this knowledge to manipulate people) as an ethical red line (see 2.5.9 on guarding against misuse).

2.3.9 Fail patterns that would weaken the role of markers

What evidence would show that somatic markers aren’t as important to the fairness story as we think? Several possibilities:

- Targeted disruption: Lesion or TMS studies find that perturbing the vmPFC, insula, or ACC (the key marker hubs) has no effect on marker-driven choices. If true, that would mean the bodily cues aren’t actually influencing decisions in a necessary way.
- No predictive power: Physiological marker signatures (HRV drops, EDA spikes, etc.) show no correlation with subsequent approach/avoid decisions under uncertainty, failing to predict behavior at all.
- Horizon indifference: In end-of-life settings or lab tasks simulating a “short horizon,” we find no interaction between horizon cues and marker strength. (In other words, people near the end feel just the same gut pulls as anyone else, with no uptick in compensatory urges.)
- Explained away: When we statistically model decisions with all known factors (utility, conflict, general arousal, learning effects), there is no QS-residual, nothing left that could be attributed to markers guiding choices toward balance. If adding learned-marker variables or Φ (feasibility-of-compensation scores) doesn’t improve predictions at all, then markers might just be redundant with ordinary decision processes.

If these patterns hold across adequately powered studies with good methods, then markers matter less than we think. We would have to revise the QS story (Chapter 5), because one of its key “felt” mechanisms wouldn’t be pulling its weight.

2.3.10 The takeaway

Somatic markers are how the body votes in your next decision. They are memory made physical, future made tangible in the present. For a law that claims life-level balance without miracles, markers are not a side note; they are one of the main channels through which balance could emerge, quietly, lawfully, one felt nudge at a time, long before we’re aware of it.

2.3.11 Where we go next:

Markers explain how yesterday’s entries tug on today’s options. In 2.4 we ground those tugs in biology, a one page tour of the body–brain loop that makes feelings lawful enough to track and perturb, tying markers to the circuits we will later test.

2.4 Affective Neuroscience in One Page

Feelings are not fog floating in some metaphysical space. They are patterns in a body-brain loop that we can describe, perturb, and measure carefully. This is a one-page tour to ground the Law of Fairness in biology without drowning in anatomy.

2.4.1 The body-brain loop (the frame)

Your nervous system constantly samples your internal state (heart, lungs, gut, temperature, muscles). The brain maps these signals, predicts the next inputs, and compares predictions with reality. Feelings arise as a continuous summary: how well are things going for me right now? Those feelings, in turn, bias actions and recruit regulatory responses (autonomic shifts, hormones, facial expressions), which change the body and create new feedback.

This closed loop, interoception → mapping → feeling → action/regulation → changed state → new interoception, gives feelings their lawful structure and makes them measurable. It is not magic; it is control. Disrupt one part and the rest changes in predictable ways (Carver & Scheier, 1990). In predictive processing terms, the brain minimizes variational free energy, a tractable bound on surprise, by updating its generative model in ways that support homeostasis (Friston, 2010).

Classic alliesthesia makes the state-link explicit: the same stimulus can flip hedonic sign with internal state. Water is intensely pleasant when dehydrated and can become aversive when overhydrated; sweet food delights when hungry but cloying when sated (Cabanac, 1971). These state-dependent reversals act like micro-guardrails against unbounded excess, consistent with our balance framing.

2.4.2 The core cast (regions and roles)

Several brain regions play starring roles in emotion and feeling:

- Insula (especially anterior insula): A hub for interoceptive maps. It integrates body signals with context to produce the felt sense that “something is happening in my body, to me.” Damping or lesions can flatten emotional experience. Hyperactivity is often observed in anxiety and pain, a kind of over-salience of bodily discomfort.
- Anterior cingulate cortex (ACC): Monitors conflict, difficulty, and error likelihood. It helps generate the urge to act or adjust. The ACC often increases activity during physical or social pain, during effort, and during impulse withholding. It is a key player whenever feeling says “this needs intervention.”
- Right inferior frontal gyrus (rIFG): Part of the brain’s braking system. It supports stopping behavior in tasks such as stop-signal paradigms. Aron (2004) provided

converging evidence for rIFG's causal role: perturbing rIFG with transcranial magnetic stimulation prolongs stop-signal reaction time, implying necessity for inhibitory control. If a strong negative marker says "don't," rIFG helps implement the stop.

- Ventromedial/Orbitofrontal prefrontal cortex (vmPFC/OFC): Central to valuation and to integrating emotion into decision-making. For converging evidence that medial prefrontal regions track value tradeoffs, Juechems (2019) reported that rACC/vmPFC activity tracks disparities between competing goals and the degree of "redress" a choice provides, implicating a network that detects and corrects imbalances in value. Damage in this territory can leave someone able to recite consequences without feeling their weight, a recipe for poor choices.
- Striatum (ventral > dorsal): Involved in reinforcement learning and habit formation. Ventral striatum (including nucleus accumbens) responds to reward prediction error ("better or worse than expected?"), classically linked to dopaminergic signaling. Schultz (1997) showed that dopaminergic neurons shift from firing to unpredicted rewards early in learning to firing to reward-predicting cues after learning, matching the error term required for reinforcement learning. Over time, striatal learning shapes which behaviors feel "worth it."
- Amygdala and Hippocampus: The amygdala flags salience (especially threat and urgency) while the hippocampus encodes context. Together they bind feelings to memory (what happened, where and when) so that similar contexts later evoke gut-level reactions.
- Default mode network (midline hubs): Including posterior cingulate and medial prefrontal cortex, active in self-referential thought and autobiographical memory. These networks sit atop deeper affective architecture identified by Panksepp (2005): mammals share primitive emotion systems (e.g., seeking, fear, care) with conserved neural substrates. This underscores that basic hedonic and aversive responses are biologically grounded across species. For example, shame involves self-modeling and social image, default-mode processing interacting with affective circuits.

2.4.3 Neuromodulators (tones that color the scene)

These are not specific locations but brain-wide messenger systems that set the tone of experience:

- Dopamine (DA): Often mislabeled the "pleasure chemical," dopamine is more centrally about motivation and learning. It signals reward prediction error and

invigorates action. High dopamine does not mean “happy”; it means “pursue and update.”

- Serotonin (5-HT): Involved in mood regulation, patience, and harm aversion. Many antidepressants modulate serotonin. It is linked to safety, contentment, and long-horizon control (“I can afford to wait”). Low 5-HT is associated with impulsivity and aggression; higher 5-HT tends to promote calm and prosocial behavior, up to a point.
- Norepinephrine (NE): Produced largely in the locus coeruleus; modulates arousal and attention. High NE supports vigilance and exploration; lower NE supports settled concentration. It functions like a gain control on alerting versus focusing.
- Endogenous opioids and endocannabinoids: Natural pain and stress modulators. Opioid release supports pleasure and pain relief (e.g., runner’s high, social warmth). Endocannabinoids support calm and appetite. Together they contribute to comfort, bonding, and “warmth.”
- Oxytocin and vasopressin: Hormones that also act in the brain and shape social bonding, trust, and empathy (oxytocin) versus protective aggression and territoriality (vasopressin). They modulate the social dimension of feeling rather than directly creating valence.

These systems modulate the body–brain loop; they do not replace it. Think of them as lighting on the stage where core circuits play out the drama of feeling.

2.4.4 Autonomic and endocrine partners (the body’s half)

On the body side, autonomic and endocrine systems matter:

- Parasympathetic (vagus nerve): The “rest and digest” system. Strong vagal tone (often indexed via higher heart rate variability) supports flexible regulation and social engagement. Slow exhalation that produces calm is a familiar vagal signature.
- Sympathetic: The “fight or flight” system. It raises heart rate, increases sweat, dilates pupils, and primes muscle tone for action. We measure it through markers such as electrodermal activity and heart rate changes.
- HPA Axis (hypothalamus–pituitary–adrenal): Governs cortisol release and longer-term stress response. Chronic stress and elevated cortisol can reshape mood and energy over hours to days. We monitor it via saliva or blood cortisol, among other proxies.

These bodily channels provide objective handles on state changes. They are often easier to measure in daily life than brain activity and they reflect aspects of feeling that matter

for our HCI measurements (e.g., anxiety tends to express as elevated sympathetic tone; contentment often tracks parasympathetic dominance).

2.4.5 Sleep and dreams (the night workshop)

Sleep is not just downtime; it is part of the affective system:

- REM sleep often carries emotional replay and transformation, a kind of theater for working through feelings. Nishida (2009) reported that REM sleep preferentially consolidates emotional memories, consistent with a role in affect integration. Scary experiences can become tamed in dreams and mundane worries can escalate into nightmares, both potentially serving processing.
- NREM sleep supports memory consolidation and synaptic downscaling, a kind of noise reduction. It can stabilize mood by pruning minor perturbations and consolidating what matters.
- Dream affect is an informative mirror of recent life. We have observed (and will examine in Chapter 10) that after especially hard days, REM dreams often carry opposite emotional themes, almost as if compensating. This “dream counterweight” phenomenon is one of LoF’s testable predictions, aligned with the broader idea that sleep can reduce the sting of painful memories.

With home EEG headbands and brief morning surveys, we can observe these processes without being intrusive. This adds another dimension to the ledger: how the night may correct or augment the day.

2.4.6 Prediction machines, not pleasure meters

A unifying insight from neuroscience (especially predictive coding) is that the brain is fundamentally a prediction engine. It continuously predicts incoming sensations and updates itself based on errors. In this view, feelings are not a bonus feature. They are downstream of prediction success or failure in relation to bodily needs and regulatory goals.

When things go as expected and needs are met, prediction error is low and we tend to feel okay or good. Pleasure (“liking”) also depends on specialized hedonic circuits, including “hotspots” in regions such as nucleus accumbens and ventral pallidum (Berridge and Kringelbach, 2015).

Joffily and Coricelli (2013) formalize valence in this framework by defining it as the negative rate of change of variational free energy over time. When free energy decreases (surprises are reduced), valence is positive; when free energy increases, valence is

negative. This ties LoF's moment-to-moment drift idea ($F(t)$) to a computational model of emotion: affect is biologically grounded, not mystical.

When things are worse than expected (or needs are unmet), errors spike and we feel bad. This helps explain why affect often oscillates around an equilibrium: larger errors drive corrections (learning and regulation) that reduce future errors. Joffily and Coricelli also show that this valence-like signal influences learning rate: when negative surprises accumulate, learning rate increases; when positive outcomes dominate, learning slows. This complements LoF's intuition that larger negative deviations tend to recruit stronger corrective actions, while positive increments tend to relax adjustment.

A crucial distinction remains: minimizing prediction error is not the same as achieving fairness. A person could minimize surprises by living in a padded room, yet remain stably miserable. The Law of Fairness concerns the integral of felt valence over a life, not merely the instantaneous reduction of surprise. Predictive processing gives us concrete hypotheses about mechanism; it does not automatically imply neutrality at closure. Brains optimize local signals; we test the long-run ledger.

2.4.7 Where the Queue System (QS) would live

If the Law of Fairness is true, QS (the hypothetical mechanism enforcing the fairness constraint) would manifest as subtle changes in known circuits, not as a new organ. Specifically, QS would appear as shifts in weights within control and valuation loops:

- We would expect rIFG and ACC (brakes and conflict monitors) to impose stronger braking and higher effort when an option threatens a non-compensable trajectory. From the inside, this can feel like: “Don’t go there. You won’t be able to pay the debt.”
- We would expect vmPFC/OFC to bias valuation in favor of options with high compensability given the current horizon. The insula may mark these valuations with gut-level urgency or relief.

What makes this testable is the concept of a QS-residual. In Chapters 5 and 6 we will define it formally, but in essence: after we account for ordinary decision factors (utilities, risk aversion, conflict, arousal, habit, and so on), is there still an extra bias in behavior that aligns with “this increases the probability my life closes neutral” versus “this makes closure implausible”? If yes, that supports QS-like influence. If not, the Law may be false or QS may not operate as proposed. We mention this here only to connect biology to theory: the candidates for QS implementation are the circuits listed above, acting in concert. Chapter 5 outlines how we would detect this experimentally.

2.4.8 What we can measure today (HCI channels)

When we build the Hedonic Composite Index, we draw on all these pieces:

- Self-report: Brief valence and arousal sliders, plus separate “gut leaning” ratings (toward/away) to capture somatic markers explicitly.
- Behavior: Reaction times to positive versus negative stimuli, approach/avoid choices, persistence on effortful tasks, and eye-gaze patterns (e.g., dwell time on negative images versus avoidance).
- Physiology: Heart rate variability (vagal tone), electrodermal activity, breathing rate, pupil size, and facial EMG, sampled by wearable sensors in daily life.
- Neural: EEG markers in lab sessions (e.g., frontal midline theta for control effort; late positive potential for emotional salience). Occasional fMRI in subsets can help calibrate EEG patterns to specific regions (e.g., linking markers to ACC engagement).
- Sleep and dreams: Sleep staging (wearable or headband EEG) plus morning dream affect reports to test counterweight patterns.

These channels feed into HCI with explicit uncertainties. None is perfect, but the composite is stronger than any single meter. Seeing measures together gives a multidimensional view of lived affect.

2.4.9 One page, one promise

Affective neuroscience does not collapse lives to circuits or chemicals. It provides a stable scaffold for measurement. It tells us where to look for measurable signals (dream counterweights, horizon effects, QS-residuals) and how to build a meter that respects the person it measures. With this scaffold in place, we can now ask the only question that matters for a law of fairness: what does the ledger say at the end of a conscious life? By understanding the biology, we ensure that we are not chasing ghosts but capturing something real.

2.4.10 Where we go next:

A sketch of circuitry is only useful if the work stays humane. In 2.5 we draw hard red lines for this project: relief before data, privacy by design, and no harm to “prove” balance, so that our methods never outrun our ethics.

2.5 What We Will Never Do

A book about fairness that measures feeling must begin with red lines. These are not marketing slogans. They are operating rules. If any are broken, the project loses its right to exist. We hold ourselves to these standards to ensure that studying the Law of Fairness never becomes an excuse to harm, intrude on, or mislead anyone.

2.5.1 Dignity over data, always

Relief is a systems variable. Comfort and dignity override data collection. In practice, that means:

- No intrusions on comfort or care. End-of-life observations will never alter clinical decisions, analgesia (pain relief), or family access. If there is any conflict between collecting a measurement and a person's comfort, dignity beats data every time.
- No instrumentalizing people. Participants are never treated as mere means to an end. Participation is optional, pausable, and consequence-free. Anyone can opt out at any point without penalty and without explanation.

2.5.2 No harm induction to “prove” balance

We will not create suffering to test whether it balances out. No trauma induction, no coercive stressors, no learned helplessness paradigms. Deliberately causing intense pain or terror just to see if relief follows is ethically repugnant. Likewise, no deprivation paradigms that withhold standard care, sleep, pain relief, or social contact. We will not deny someone something they need in the name of research.

- No aversive conditioning to implant markers. We will not attempt to “wire in” new fear or shame responses through repeated shocks or humiliation..
- No wireheading. We will not do the opposite either: no direct brain stimulation to induce artificial euphoria or numbness to manipulate the ledger. If the Law only held under cruel or artificial intervention, it would not be worth holding.

2.5.3 End-of-life ethics are non-negotiable

When studying people near the end of life (where the ledger closes), we follow strict rules:

- Comfort first. Morphine before measurement, always. If a person is in pain, alleviating it takes absolute priority over collecting any data point.
- Non-interference. We observe what naturally occurs. We do not stage or encourage “Hollywood moments.” No orchestrating reconciliations, confessions,

or farewell scenes. If those happen, they are the person's choice or the family's, not ours.

- Consent and assent. If someone is conscious and able, we get informed consent. If they are not, we rely on ethically approved surrogate consent and bedside assent (signals of comfort or willingness). The person can always signal "no" through a gesture or expression, and we stop.
- Right to silence. Anyone can withdraw at any time without a reason. If a dying person does not want to answer another question or wear a sensor, that wish is honored immediately. Participation is a gift, not an obligation.

2.5.4 Animal research: strict minimization

Where we include animal models (for basic science on circuits), we follow the "3 R's": Replace, Reduce, Refine. We replace animal tests with simulations or existing data whenever possible. We reduce the number of animals to the minimum needed for statistical validity. We refine procedures to eliminate or minimize suffering, preferring noninvasive observation in comfortable settings over invasive experiments.

Prohibited paradigms: We explicitly rule out harm-centered studies that some might imagine as "tests of balance" (e.g., learned helplessness-style protocols). If discussed, it is only as thought experiments or references to existing literature, never as new work endorsed by this project.

2.5.5 Simulation ethics: non-sentience guarantees

We use computational simulations to test parts of the theory (for example, simulating many lifetimes under assumptions and observing emergent patterns). However:

- No sentient agents, ever. We simulate only agents that are complex algorithms with no capacity for experience or suffering, no AI that even approaches conscious experience. If an AI could feel, it would fall under human-subject-style protections, which would rule out uncontrolled experiments. (If the simulation hypothesis is real, we predict a creator would implement protections such as LoF for sentient agents.)
- Red-team audits. External experts review models to ensure we are not inadvertently creating agents that suffer. We maintain kill-switch protocols if any simulation unexpectedly showed signs of agency or distress.
- No dark patterns. Simulations are not connected to real users in manipulative ways. We are not secretly tweaking live systems to see whether we can push emotions around. Simulation stays in simulation and any use of user data is opt-in and transparent.

2.5.6 Data dignity and privacy by design

Data about feelings is deeply personal. We design our pipeline to respect that:

- Minimize and silo: We collect only what we need for preregistered analyses. Data are encrypted at rest (device and server) and in transit. Identifiers are stored separately from sensor and survey data, with coded linkage, so a single breach is less likely to expose identity.
- Participant control: Participants can view their own data, correct obvious errors, or request deletion. They can choose the scope of sharing (aggregate-only, approved secondary analysis, or no sharing) and can revoke permission later. At revocation we stop using future data and, where possible, remove past data if requested.
- Differential privacy in releases: When releasing datasets, we anonymize and add statistical noise to reduce re-identification risk and suppress uniquely identifying records.
- No commercial reuse: Data are never sold or used for marketing, credit scoring, employment decisions, or provided to law enforcement without an official warrant and ethics board review. No one should face real-world penalties from “happiness data.”

2.5.7 No p-hacking, no post-hoc endpoint drift

We touched on this in 2.2.6, but it is worth re-emphasizing as an ethical stance:

- Preregistration: Primary outcomes, equivalence bounds for neutrality, key ROIs, and stopping rules are decided before viewing outcome data. No moving goalposts.
- Transparent deviations: If changes become necessary (e.g., a sensor fails and we switch metrics), we document the change immediately with a timestamp and reason. Both the original plan and revision are reported side by side.
- Negative results are published: If predicted signals (dream counterweights, horizon scaling) fail to appear in adequately powered tests, we report that. Absence of evidence, when the test is strong, is evidence of absence for that effect.

2.5.8 Communication ethics

- No moral cover: We will never say or imply that present suffering is “okay” because a ledger might balance later. Present suffering deserves present care. The Law of Fairness, if true, is descriptive, not prescriptive. It is about what happens, not what should happen.

- Plain disclaimers: We label speculation as speculation, hypothesis as hypothesis, and results as results. We correct public misunderstandings and treat preventing metaphysical overreach as an ethical duty.
- Context shielding: Anecdotes and case studies are anonymized rigorously. We may use composites or mild obfuscations unless we have explicit permission to tell an identifiable story.

2.5.9 Guarding against dual-use

- No tools for manipulation: If we discover ways to influence emotions, we do not package them into a playbook for persuasive tech or propaganda. We focus on protective uses and treat induction methods as sensitive.
- Governance layer: An external ethics board, including community members and patient advocates, reviews releases and has veto power. If something could be misused or badly misunderstood, we do not release it without safeguards, or at all.

2.5.10 No teleology, no metaphysical smuggling

- Constraint language only: We do not write “the universe wants balance” or “fate ensures fairness.” We frame LoF as a constraint on admissible trajectories, not a goal pursued by an agent. We say “trajectories that would prevent neutral closure are pruned” rather than “nature prevents unfairness.”
- Causal closure respected: We do not invoke forces outside known physics or biology. If evidence pointed that way, we would admit we do not know what is happening rather than imply something supernatural. The moment this claim demanded supernatural explanations, we would withdraw it.

2.5.11 Respect for dissent and vulnerability

- Opt-out culture: Disagreement is welcomed. No one is pressured to “believe.” Quitting participation is always allowed.
- Protected groups: For children, cognitively impaired adults, or anyone under duress, we add safeguards and independent oversight. We exclude populations that cannot give real consent unless justified with extreme caution and ethics approvals.

2.5.12 Incidental findings and duty of care

- Clinical handoff: If monitoring suggests unmanaged pain, severe depression or suicidality, or dangerous health issues, we have protocols to alert the participant

and connect them to care, with consent whenever possible. We are not a healthcare provider, but we do not ignore serious risk.

- No unwelcome surprises: Participants choose in advance what feedback they want. The default errs toward safety for life-threatening risks; for less acute issues, we follow participant preferences.

2.5.13 Conflict of interest transparency

- Declare everything: Funding sources and relevant ties are disclosed.
- Firewalls: Funders do not control analysis or publication. Unfavorable results cannot be suppressed.

2.5.14 Stop rules that end studies early

- Burden thresholds: We predefine thresholds for participant burden or distress and monitor actively. If participation itself is harming people or producing excessive anxiety or dropout, we stop and reassess.
- Futility rules: If interim analysis indicates predicted effects are very unlikely to be detected even with more data, we consider stopping to avoid wasting time and burden.

2.5.15 The fail-fast pledge

We make a global fail-fast pledge: if robust, well-powered evidence accumulates against the Law of Fairness, we publish it, retract or revise the claim, and pivot to what the evidence suggests. We do not move goalposts.

Fail pattern (Law of Fairness overall):

- Terminal imbalance outside bounds: Clear, replicated evidence of lifetime ledgers ending significantly positive or negative *outside* the predefined neutral band for a substantial number of people *with all channels functioning*. (In plain terms, if many well-tracked lives end with strong evidence of more suffering than joy or vice versa, well beyond our \pm neutral margin.)
- Absent compensatory dynamics: Failure to observe the two key signatures of balancing in any form, no horizon scaling and no dream counterweights, under adequate power and methods.
- Null QS-residuals: Robust evidence that after accounting for known decision factors, there is no leftover pattern attributable to fairness constraints, no added explanatory power from LoF terms.
- Rival model matches all signatures: A rival theory (e.g., reinforcement learning plus homeostatic regulation without a fairness constraint) reproduces all

signatures and predicts new data as well as or better than LoF-based models in out-of-sample tests.

If any of the above are observed with strong evidence, we will say so plainly, publish prominently, and abandon or radically revise the claim. No hazy beyond. The claim lives or dies by data.

These red lines and fail criteria keep the science worthy of the people it studies. They ensure that even if we are wrong, we do no harm while finding out. With boundaries set and our ethical compass checked, we can now preview the unit and the account that make a life-level claim testable.

2.5.16 Where we go next:

With ethics fixed, we put a unit and an account on the table. In 2.6 we define Hedonic Composite Units (HCU) and the life ledger, and show how uncertainty travels all the way to the final neutrality test at the death of mind.

2.6 Preview: Hedonic Composite Units (HCU) and the Life Ledger

To preview how the ledger is estimated in practice, we calibrate the moment-to-moment net-affect rate using a delta-based Hedonic Composite Index (HCI). On a fixed sampling grid $\{t_i\}$, we compute standardized first differences from multiple channels (self-report, physiology, sleep and dream affect, behavior, neural markers) and combine them with preregistered weights:

$$HCI(t_i) = \sum_k w_k \cdot \Delta z_k(t_i)$$

Here, k indexes measurement channels and $\Delta z_k(t_i)$ denotes the standardized change in channel k at time t_i . This delta-based HCI serves as our empirical estimate of the moment-to-moment net-affect rate $F(t_i)$.

With anchors in place, the lifetime ledger estimate $\hat{L}(T)$ is the time integral of this net-affect estimate $HCI(t)$, expressed in Hedonic Composite Units (HCU).

$$\hat{L}(T) = \int_0^T HCI(t) dt \text{ (in HCU units)}$$

where $\hat{L}(T)$ is measured in HCU and represents the cumulative balance of felt experience up to time T .

Later chapters formalize the mechanism and the neutrality test in full detail. This preview links the meter (HCI) to the account (the life ledger estimate $\hat{L}(T)$) without introducing the deeper formal machinery too early.

Allostatic control. Biologically, the brain maintains stability by changing. It anticipates needs, updates set points, and deploys autonomic, hormonal, and behavioral corrections. Chronic deviation carries allostatic load, cumulative wear and tear from overusing stress systems that degrades health and fitness. In our terms, those regulatory operators act on $S(t)$, the organism's drive-load, and their felt report is $F(t)$. Neutrality, on this picture, is not mystical balance but the long-run consequence of repeated prediction and correction. Momentary states can be far from neutral; across a lifetime, the accumulated ledger can approach equilibrium.

If feelings are the final currency, we need a unit for that currency and an account to place it in. The unit lets us add moments without pretending every moment is identical. The account lets us see where a life stands in aggregate. This section provides the intuitive frame that carries through the rest of the book. The full formal structure appears later; here we establish the conceptual link between measurement and lifetime balance.

2.6.1 From many clues to one meter

Back in Section 2.2, we introduced the Hedonic Composite Index (HCI), a cautious, moment-to-moment estimate of net affect built from converging clues: self-reports, behavior, physiology, neural signals, and dream affect. HCI is a latent estimate with explicit uncertainty; it is not any single channel. You can think of HCI as the needle on a dashboard. It rises with pleasant moments and falls with distress. Its confidence band narrows when multiple channels converge and widens when data are sparse or noisy. HCI is not designed for sweeping judgments in isolation; it is designed to feed the ledger with transparent error.

2.6.2 The unit: Hedonic Composite Unit (HCU)

To speak quantitatively about feeling, we need a common unit. We define 1 HCU as one minute at a calibrated +1 level on the HCI scale. That +1 is anchored to tangible experiences. For example, +1 might correspond to a mild but noticeable uplift from music or friendly contact; -1 might correspond to a brief social slight or minor discomfort. Multiple positive and negative anchors ensure that larger magnitudes, such as +5 or -10 HCU, map to experiences of roughly comparable intensity. The detailed calibration appears in Chapter 7. Conceptually, HCU converts fluctuation into an additive unit.

If your HCI is approximately +2 for 10 minutes, that contributes +20 HCU. If later it is -1 for 30 minutes, that contributes -30 HCU. The net for that interval is -10 HCU, with uncertainty carried from each estimate. Calibration proceeds within person first and is then linked across people, so that +10 HCU represents a similar magnitude of felt change across individuals within stated error bounds.

2.6.3 The account: the Life Ledger L(T)

Let $F(t)$ denote instantaneous net affect at time t . The life ledger up to the end of the stream at time T is the time integral of that affect:

$$L(T) = \int_0^T F(t) dt \text{ (in HCU units)}$$

In plain terms, we sum all positive and negative contributions across the entire life. Each moment contributes its increment. Because $F(t)$ is not directly observable, we estimate this quantity empirically via $\hat{L}(T)$ as defined above. Uncertainty is carried forward through this integration. At any point, $L(t)$ has a confidence interval reflecting accumulated measurement error. We can summarize daily net HCU with a 95 percent interval, then aggregate across months and years, and finally estimate $\hat{L}(T)$ at closure with its uncertainty.

The Law of Fairness predicts that, for each unified conscious stream, the final confidence interval for $\bar{L}(T)$ lies within a predefined neutrality band around zero. We do not require the point estimate to equal zero exactly. We require that, after accounting for uncertainty, it cannot be distinguished from neutral beyond the narrow band specified in advance.

2.6.4 Equivalence bounds: “neutral” has a ruler

Neutrality is operational, not poetic. Before collecting outcome data, we define a small band around zero in HCU units that counts as practically neutral. Let $\pm K$ HCU denote that band. At closure, each ledger is classified as follows:

- If the entire 95 percent confidence interval of $\bar{L}(T)$ lies within $\pm K$ HCU, the ledger is Neutral (within bounds).
- If the 95 percent interval lies entirely above $+K$ or entirely below $-K$, the ledger is Non-neutral.
- If the interval overlaps the boundary, the result is Inconclusive.

This prevents impressionistic judgments. Balance must fall within a band specified in advance. Conversely, if a ledger clearly ends outside that band with sufficient power and intact measurement, the law faces a direct challenge.

2.6.5 Why dreams and horizons matter to the ledger

Two signatures are especially relevant to how the ledger evolves over time: dreams and time horizons.

- Dream counterweights. REM dreams often shift the emotional tone of the preceding day. After sequences of negative affect, dream content may tilt positive more often than chance would predict. The Law treats this as a measurable pattern. If a ledger leans negative, nights may provide small upward corrections beyond ordinary adaptation. Dream reports and sleep-stage data allow this to be tested directly.
- Horizon scaling. As perceived remaining time shrinks, compensatory experiences are expected to intensify. Nearing the end of life, the organism may seek or become more receptive to experiences that move the ledger toward neutrality. Measurably, this implies larger-magnitude HCU episodes under short horizons, conditional on intact channels for compensation. Hospice studies and laboratory tasks simulating time scarcity provide ways to test this prediction.
- Both signatures are visible in the ledger record. We can examine whether dream affect after negative days adds positive HCU, and whether end-of-life periods show structured changes linked to prior imbalance.

2.6.6 How uncertainty stays honest

Uncertainty is propagated at every stage, from instantaneous HCl to daily HCU to the lifetime $L(T)$. If HCl is estimated from sparse or noisy data at a given moment, its variance increases. When integrating across time, that variance is incorporated rather than ignored. The neutrality test is applied to the confidence interval of $L(T)$, not merely to a point estimate.

For example, if a final 95 percent interval for $L(T)$ is $[-5, +4]$ HCU and the neutral band is ± 3 HCU, the case is inconclusive. We cannot claim neutrality or non-neutrality with confidence. This discipline prevents overstatement when measurement is imprecise.

2.6.7 A one-day example (intuitive)

Consider a simplified day converted to HCU:

- Morning run: HCl at $+1.5$ for 20 minutes $\rightarrow +30$ HCU
- Commute stress: HCl at -0.8 for 40 minutes $\rightarrow -32$ HCU
- Lunch with a friend: HCl at $+0.7$ for 50 minutes $\rightarrow +35$ HCU
- Afternoon fatigue: HCl at -0.5 for 60 minutes $\rightarrow -30$ HCU
- Evening music practice: HCl at $+1.2$ for 30 minutes $\rightarrow +36$ HCU
- Late-night worry: HCl at -0.6 for 25 minutes $\rightarrow -15$ HCU

End-of-day net = $+24$ HCU, with uncertainty (for example ± 8 HCU). A single day reveals little. Thousands of days, aggregated with their uncertainty, form the ledger. The example illustrates how positive and negative periods add and subtract within a consistent unit.

2.6.8 When the meter would fail us (and what that means)

The ledger depends on the integrity of its meter. Several failure patterns would undermine it:

- Unstable calibration. If the mapping from composite indicators to HCl drifts unpredictably over months or years, long-run integration becomes unreliable.
- Persistent cross-cultural non-invariance. If HCU values cannot be made comparable across contexts despite careful anchoring and invariance testing, a unified ledger becomes questionable.
- Ever-widening uncertainty. If accumulated uncertainty on $L(T)$ remains so large that neutrality can never be meaningfully tested, the claim becomes empirically out of reach.

In these cases, the issue would not automatically falsify the Law of Fairness. It would indicate that the measurement framework is insufficient. The appropriate response

would be to publish the limitation and reassess the empirical program rather than press forward with unreliable tools.

2.6.9 The preview's promise

From this point forward, the book treats HCU as the calibrated unit of felt change and $L(T)$ as the cumulative ledger of a unified conscious stream. Neutrality at closure is defined by explicit bounds and uncertainty is never erased. With a unit and an account in place, the life-level claim moves from metaphor to measurement. It becomes a statement that can be examined in data and, if necessary, relinquished if the evidence demands it.

2.6.10 Where we go next:

The meter, the unit, and the ledger are now in view. Part II states the Law of Fairness plainly, one sentence and fully defined, and names the signatures that must appear, and the failures that would end the claim, if a lifetime balance is real.

Part II — The Law, Stated Clearly

Part I asked the question of fairness and set the stakes in human terms. Part II states the Law in language tight enough to test and clear enough to read. We define every term the claim depends on, give precise expressions that make it auditable, and spell out exactly what observations would count against it. No metaphors, no averages, no hope as evidence. A claim aspiring to lawhood must be crisp, formally stated, and falsifiable. Part II begins that audit.

We adopt a state-change formalism. Let $L(t)$ denote the latent life ledger and $F(t)$ the instantaneous net-valence rate of a unified conscious stream. The state equation is:

$$dL/dt = F(t).$$

Empirically, we estimate the ledger as:

$$\hat{L}(t) = \int_0^t HCl(\tau) d\tau,$$

where HCl is a delta-based composite of first differences across self-report, physiology, neural, behavioral, and dream channels, combined with reliability weights and preregistered invariance checks. In discrete implementation, $\hat{L}(t_n) = \sum_{i=1}^n HCl(t_i) \cdot \Delta t_i$.

The theoretical closure condition is:

$$L(T) = \int_0^T F(t) dt = 0$$

This equality is the law's boundary condition at the death of mind. It asserts a strict closure guarantee for each unified conscious stream.

Because $F(t)$ is not directly observed, closure is tested operationally using equivalence bounds $\pm K$ (in HCU) applied to the measured ledger $\hat{L}(T)$ under a preregistered uncertainty model. Neutrality is adjudicated by equivalence testing within $\pm K$; this separates the exact theoretical target $L(T) = 0$ from its empirical estimate $\hat{L}(T)$, specifies how evidence accumulates, and makes failure conditions explicit.

We treat fairness as a candidate natural law, held to the same standards as any physical law: unambiguous definitions, formal metrics, and concrete break conditions. Because the subject is conscious experience, the burden of proof is high. Part II specifies who the law applies to (each unified conscious stream), what “neutral” means (a defined zero balance within K , not moral justice), and why—if it exists—such balance would be a constraint of nature rather than wishful thinking.

By the end of this Part, we place a one-sentence law on the table and provide a framework for checking it. That framework lists explicit fail patterns, including: a well-measured life

with $|L(T)| > K$ after all preregistered error budgets; persistent divergence between $\hat{L}(T)$ and $L(T)$ not explained by model diagnostics; absence of predicted end-of-life compression when compensatory channels remain open; and identity misassignment not governed by split, merge, and pause rules. These chapters do not prove the Law. They make it specific enough that any competent team could attempt to break it. If the Law holds, it functions as a constraint on allowable life trajectories, not as a guiding purpose.

In short: Part II turns fairness into a testable proposition. Net valence is treated as a state variable whose rate $F(t)$ integrates to a lifetime ledger $L(T)$. The law's boundary condition is $L(T) = 0$ at the death of mind; operationally, we test the estimate $\hat{L}(T)$ against preregistered equivalence bounds $\pm K$. We treat the claim analogously to a conservation law: assign units, write the accumulation equations, define the boundary condition, and preregister what would count as failure. If a well-measured stream ends with $\hat{L}(T)$ outside $\pm K$ after all preregistered error budgets, the idea is wrong. Part II invites that decisive test by stating the Law plainly and highlighting how it could fail.

What this Part will do for you:

- Distill the Law in one sentence, with full definitions.
Law of Fairness (one sentence): For each unified conscious stream, from the onset of sustained consciousness to the death of mind at time T , the lifetime integral of net valence lies within a preregistered neutral band ε around zero—that is, $L(T) = \int_0^T F(t) dt$ with $|L(T)| \leq \varepsilon$ —where “unified conscious stream,” “death of mind,” and “neutral band” are defined in operational, testable terms.
- Show the formal audit trail. We write three linked relations: (i) the state equation $dL/dt = F(t)$; (ii) the measured ledger $\hat{L}(t) = \int_0^t HCl\Delta(\tau) d\tau$; and (iii) the closure condition $|L(T)| \leq \varepsilon$ assessed via preregistered equivalence tests. We specify error budgets, invariance checks across groups, and model diagnostics so that every inference about neutrality is numerically auditable.
- Draw a bright line between a true law vs. mere tendency. Phenomena such as hedonic adaptation, opponent-process rebounds, or treadmill effects describe regression toward a baseline mean. They do not entail lifetime ledger closure. We state predictions that differ from those tendencies, including approach to the neutral band at closure for each stream, and we specify the observations that would contradict the Law even if adaptation holds on average.
- Declare all boundary conditions up front. We preregister identity rules: one ledger per unified stream, split and merge criteria (e.g., split-brain or dissociative cases), and pause rules for dreamless sleep, coma, and surgical anesthesia. We define scope (which lawful minds count) and minimal duration. These rules prevent post

hoc redefinition of “self” or “experience” to rescue the theory and ensure experiments remain interpretable.

- Frame the Law as a passive constraint, not a purpose. Neutrality is treated as a conservation-like restriction on allowable life trajectories, not as a goal pursued by brains or by nature. No teleology, no moral desert, and no license to engineer suffering. The claim is descriptive and stands or falls on measurement.
- Identify fail patterns and testable predictions. We enumerate disconfirmations such as: well-measured ledgers ending outside ε ; systematic group-level biases in $\bar{L}(T)$ persisting after invariance correction; absence of predicted end-of-life compression with compensatory channels intact; or failure of dream- and narrative-mediated counterweights to contribute measurable Δ toward neutrality under prospective modeling.
- Carry forward an ethical safeguard. Throughout, we uphold the principle that relief is a systems variable. Comfort and dignity override data collection. No matter how intriguing the hypothesis, we never withhold analgesia or care for measurement. That safeguard travels into the formal work, ensuring rigor never outruns compassion.

Chapters in this Part:

- **Chapter 3 — The Law of Fairness** - States the claim in one sentence, unpacks each term (unified stream, ledger, neutral, death of mind), and lays out six assertions that render the law testable. It draws boundaries around what the Law does not say and sets the evaluation endpoint at the death of mind, not legal or cardiac death. Research notes specify the ledger integral and criteria for neutrality.
- **Chapter 4 — Constraint, Not Purpose** - Clarifies what kind of thing the Law would be if true: a constraint on outcomes, not a cosmic goal or moral mandate. It contrasts constraint-based explanations with teleological stories and situates the claim among accounts of natural laws. Research notes address regularity checks and statistical safeguards needed to ensure the Law is properly tested.

Where we go next:

We now move to Chapter 3, where the Law is stated, translated into plain speech, and bounded so it can be checked. Keep in mind the refrain from Part I: relief is a systems variable; comfort and dignity override data collection. That safeguard remains intact as we move into formal ground.

Chapter 3 — The Law of Fairness

A young researcher faces a skeptical room. On the whiteboard is a single sentence: “Each unified conscious stream ends with a neutral ledger of felt experience at the death of mind.” Some nod; others frown. If this is a law of nature, it must be defined precisely—and made falsifiable.

To convince a critical audience, we must specify exactly what the Law says, translate it into measurable quantities, and identify experiments that could refute it. That is the task of this chapter.

By the end of this chapter, you will know what the Law claims, what it does not claim, what conditions must hold for it to be true, and what evidence would count decisively against it. Fairness is attached to a single bearer (the unified conscious stream), a single currency (the cumulative balance of pleasant versus unpleasant experience), and a single timeframe (the end of that stream, the death of mind). We present the Law in plain English and in compact formal form so that supporters and critics alike argue about the same proposition.

First, we introduce the Law in one sentence and unpack each term. Next, we lay out six core assertions: conservation at closure, universality, neutral termination, fairness as constraint rather than intent, resolution of extreme suffering, and a candidate mechanism enforcing it. Each assertion carries measurable signatures and explicit fail conditions. We then draw boundaries around what the Law does not say: no cosmic intent, no moral desert, no excuse to ignore suffering, no appeal to afterlife or karma, no averaging across persons. Fairness here is not justice; it is a proposed balancing constraint.

We then fix all boundary conditions. We define what counts as a continuous stream, how ledgers pause or resume (sleep, anesthesia), how identity splits or merges, how extremely short lives are treated, and how non-human minds are considered. These definitions are set before examining outcomes so that edge cases cannot be adjusted post hoc.

With terms fixed, we present the formal equations for the lifetime ledger and the neutrality test at closure, explaining why the stopping point is the death of mind. We describe predicted signatures as closure approaches, such as drift toward the neutral band and compression of ledger variance across streams. A research notes section provides the technical backbone: estimation of $F(t)$ via HCl, integration into $\bar{L}(t)$ with propagated uncertainty, handling of branching or fusion, and statistical equivalence testing at closure.

What you'll get from this Chapter:

- A claim you can hold in one sentence. By the end of this chapter, you will be able to state the Law of Fairness from memory, know exactly what each word means, and understand how that sentence translates into a measurable proposition. Nothing rests on metaphor or intuition; every term is fixed.
- A structure that can be broken. You will see the Law decomposed into a small set of explicit commitments. Each commitment generates observable signatures and corresponding failure modes. If any one of them collapses under adequate measurement, the larger claim collapses with it.
- Clean separation between science and story. You will see precisely where the Law stops. It makes no appeal to cosmic intent, moral desert, reward, punishment, afterlife, karma, or averages across people. It is framed strictly as a constraint on trajectories of experience, not a narrative about justice.
- Rules declared before data. You will see the identity criteria, pause rules, scope conditions, and endpoint definitions fixed in advance. These boundary conditions prevent retrospective adjustment of what counts as a “life,” a “self,” or a “moment” to protect the theory.
- A neutrality test with teeth. You will understand how neutrality is operationalized using a preregistered equivalence band $\pm K$ HCU, why closure is evaluated at the death of mind, and what empirical pattern would decisively count against the Law. If lives end outside ϵ under intact measurement, the claim fails.
- The technical backbone in plain view. You will see how the latent affect process $F(t)$ is estimated as HCl, how that signal is integrated into a running ledger $L(t)$ with uncertainty, and how equivalence testing at closure is conducted. The equations are not decorative; they are the audit trail.

Subsections in this Chapter:

- **3.1 Canonical Statement** – Introduces the one-sentence Law and defines each term in operational form so that no ambiguity remains.
- **3.2 Six Assertions of the Law** – Specifies the commitments implied by the canonical sentence, along with their predicted signatures and explicit fail conditions.
- **3.3 What the Law Does Not Say** – Draws hard boundaries around teleology, moral narratives, and misuse, keeping the claim descriptive and constraint-based.
- **3.4 Boundary Conditions** – Declares identity rules, pause criteria, and scope limits in advance so ledgers are not reassigned or redefined post hoc.

- **3.5 The Death of Mind** – Defines the closure point clinically and ethically, sets stabilization criteria, and explains how neutrality is evaluated at endpoint without compromising care.
- **3.6 Research Notes: The Ledger Integral** – Provides the formal relations $dL/dt = F(t)$, $\dot{L}(t) = \int_0^t HCl(\tau) d\tau$, and $L(T) = \int_0^T F(t) dt$, along with uncertainty propagation and equivalence testing procedures.

Where we go next:

We start with 3.1. The single sentence is the handle you'll use for the rest of the chapter. Each subsequent subsection simply widens that grip until the test is ready to run.

3.1 Canonical Statement

Every unified conscious stream ends with a neutral ledger of felt experience at the death of mind.

3.1.1 Unpacking each word

Unified conscious stream: The “owner” of the ledger is the single experiencing subject that can access information, integrate it, and guide action as one continuous identity over time. In simpler terms, it is the one consciousness, the entity that can say “I,” that persists through a life. If that unity is absent, for example during deep general anesthesia or dreamless sleep, ledger accrual stops because there is no unified experience occurring. If unity splits (rare cases such as split-brain patients or severe dissociative episodes), separate ledgers branch under fixed rules (see Section 3.4). If unity later returns (streams fuse), those ledgers rejoin by simple summation. In short, the Law pertains only to coherent conscious streams. If there is no stream, there is no ledger accruing.

Ledger of felt experience: This is the running total of all pleasant and unpleasant experience in that stream’s life. We construct it by integrating net affect over time. In practice, we estimate net affect moment to moment with the Hedonic Composite Index (HCI, introduced in Section 2.6 and detailed in Chapters 7–8) and accumulate it in Hedonic Composite Units (HCU). The ledger always carries uncertainty. Each increment has an error margin that propagates forward, so we never treat the ledger as perfectly known. It is a probabilistic sum of what that life felt like, continuously updated as the person lives.

Neutral: Neutral is not a fuzzy feeling or a metaphor. Neutral means that, by the time the stream ends, the lifetime ledger’s uncertainty interval lies entirely inside a small preregistered equivalence band around zero. Within the resolution of our measurement, total pain and total pleasure are indistinguishable from zero. If the final uncertainty interval lies entirely outside the band, the Law is violated for that stream. If the interval overlaps the band boundary, the outcome is inconclusive: we cannot classify it as neutral or non-neutral under the preregistered criterion. “Neutral” is therefore a strict, falsifiable condition, not an interpretive one.

Death of mind: The point at which the conscious stream permanently stops experiencing anything. It is not necessarily the last heartbeat or the legal time of death. It could occur earlier than physical death (if the brain irreversibly loses unified consciousness before cardiac arrest), or slightly later (if brief resuscitation or lingering brain activity sustains conscious experience after clinical death). We operationalize it as the last moment T at

which the Unity Index (our measure of unified consciousness) is above threshold (see Section 3.5). After that point, the stream has ended and the ledger closes.

3.1.2 Plain-language gloss

In plain language, the Law of Fairness says: when a life's conscious experience comes to its end, the total sum of how it felt, all the happiness and all the suffering, balances out to neutral. This is not because the universe is kind or because people always make wise choices. Rather, if the Law holds, it is because only those life trajectories are possible in a world constrained by this law.

The Law does not promise gentle days or equal fortunes. It does not claim that resources, opportunities, or outcomes are fair. It makes a narrower claim about life histories: paths that would leave a conscious stream with an uncompensated surplus of either pain or pleasure by its end are eliminated. Either they do not occur, or if they begin to occur, they are redirected toward neutral through ordinary mechanisms. We call the hypothesized enforcement of that constraint the Queue System (QS), introduced in outline here and expanded in Section 3.6 and later chapters.

3.1.3 What would make the sentence false

To keep the claim testable, we declare up front what would falsify the one-sentence statement. Clear failure scenarios include:

- A well-identified stream that ends outside the band: If we find even one convincingly measured life whose final ledger uncertainty interval lies entirely beyond the neutral equivalence bounds (beyond $\pm K$ HCU, where K is the preregistered margin), that violates the claim for that life. One clear counterexample is enough to call universality into question.
- No horizon effects or dream counterweights in compensable situations: If people nearing end of life with comfort channels open show none of the predicted opposite-valence dynamics (no horizon-linked compensations), and if dream affect never reflects prior-day imbalances beyond chance when dream measurement is intact, then the compensatory signatures the Law predicts are absent. In simpler terms, if we repeatedly fail to observe telltale balancing dynamics where they should be most detectable, the Law loses support.
- No QS-residuals in control hubs: If rigorous analyses of decision-making and neural activity find no trace of the predicted QS bias, that is, no residual effect aligned with ledger imbalance or remaining horizon after accounting for standard drivers of behavior, then the Law lacks its proposed mechanism. QS is posited to

influence choice weights and control signals subtly; if comprehensive modeling finds no such influence, plausibility drops.

- Ledger variance that grows toward the end: If longitudinal data show that lifetime ledger totals become more spread out as death approaches, or fail to compress at all, then there is no evidence of a converging constraint. The Law predicts convergence toward neutrality near closure. Systematic divergence would contradict that.
- A rival theory explains it all without a law: If another model built from known processes can reproduce neutral end-ledgers and all predicted signatures without a global fairness constraint, then the Law is unnecessary. If it adds no explanatory power beyond ordinary mechanisms, it should be set aside.

3.1.4 Research note (formal definitions)

We now formalize the ledger and neutrality criterion. Let $F(t)$ denote the latent net affect rate, the true (unobservable) momentary hedonic state, for a unified stream over its conscious lifespan $[0, T]$. Define the true life ledger as the time integral of this process:

$$L(T) = \int_0^T F(t) dt \text{ (measured in HCU).}$$

Because $F(t)$ is not directly observed, we estimate it with the Hedonic Composite Index, $HCI(t)$, and compute the estimated running ledger:

$$\hat{L}(t) = \int_0^t HCI(\tau) d\tau$$

By the terminal time T (death of mind), we obtain a final estimate $\hat{L}(T)$ with propagated uncertainty. Under the Law of Fairness, a unified stream must satisfy neutral closure at T . We operationalize this as an equivalence test on the lifetime ledger using preregistered bounds $\pm K$ HCU (K is the allowed neutral margin):

- Neutral: The 95 percent credible interval (Bayesian) or equivalence-test confidence interval (frequentist) for $\hat{L}(T)$ lies entirely within $(-K, +K)$.
- Non-neutral: The interval lies entirely above $+K$ or entirely below $-K$.
- Inconclusive: The interval overlaps the boundary of $(-K, +K)$, so the result is indeterminate under the preregistered criterion.

In a Bayesian analysis, we may declare neutrality if $P(-K < L(T) < K | \text{data}) \geq 0.95$. In a frequentist Two One-Sided Tests (TOST) approach, we require that the 90 percent confidence interval for $\hat{L}(T)$ lies entirely within $(-K, +K)$, corresponding to $\alpha = 0.05$. Neutrality as equivalence: We adjudicate neutrality via TOST by preregistering equivalence bounds $[-K, +K]$ (the SESOI). At T , reject H_0^- : $L(T) \leq -K$ and H_0^+ : $L(T) \geq +K$. Equivalently, the 90 percent confidence interval for $\hat{L}(T)$ lies entirely within $(-K, +K)$. We

report K justification, intervals for $\bar{L}(T)$ and $L(T)$ where modeled, and sensitivity analyses for missingness and sedation. This makes “neutral” auditable, not rhetorical.

An identity criterion underpins these tests. All of this assumes we correctly identify the unified stream across the life. In difficult cases (e.g., prolonged minimally conscious state, resuscitation), we use the Unity Index (detailed in Section 3.5) to adjudicate whether we still have the same continuous conscious stream or whether it ended earlier. This ensures the bearer of the ledger is defined consistently and we do not integrate the wrong intervals.

Mechanistic predictions (for those interested): The formal model for how neutrality might be enforced (via QS) yields quantitative signatures. A fuller list appears later (and in Section 3.6), but examples include:

1. Horizon-weighted selection: As subjective remaining time H shrinks, the distribution of available actions $A(t; \bar{L}, H, C)$ shifts, narrowing and reweighting toward options with higher feasibility-of-compensation Φ . Operationally, as time grows short or the ledger skews, choices should tilt toward restorative actions. The menu of life converges toward options that keep closure attainable.
2. QS-residuals in control networks: In control regions (e.g., rIFG, ACC, vmPFC), after accounting for standard predictors of choice (utility/benefit U , conflict C , arousal, motivational salience), a residual effect should correlate with Φ . If the Law holds in the proposed form, we expect this compensability term to show a positive association: higher Φ biases both control-hub activity and policy selection toward repair options.
3. Variance compression: The cross-sectional variance of $\bar{L}(t)$ across individuals declines as $t \rightarrow T$. We detect this by tracking cumulative ledgers over time. If the Law holds, the spread of ledgers narrows near the end. (Functional data methods or variance-trend tests can evaluate whether $\text{Var}[\bar{L}(t)]$ decreases as t approaches T .)
4. Dream-phase inversions: Models predicting dream affect from prior-day imbalance should show an elevated inversion effect: dream affect tends to counteract the preceding day’s net affect. Strong negative imbalance before sleep predicts a higher probability of positively toned dreams (and vice versa). We test this in dream reports and REM measures (see Chapter 10).

We will detail these predictions later. The point here is that the formalism does not merely state the Law. It generates concrete byproducts that can fail.

3.1.5 Where we go next:

With terms fixed and the endpoint clear, we now turn to what must be true if the sentence is true. Section 3.2 names six assertions and the signatures each would leave in real data. Those assertions are the scaffolding for the tests that follow.

3.2 Six Assertions of the Law

The Law of Fairness rests on six key assertions. Each assertion articulates a necessary part of the hypothesis, why it matters, and how we would know if it failed. Together, these assertions form the structure supporting the Law. If any one repeatedly fails under robust evidence, the Law as a whole must be reconsidered or abandoned.

3.2.1 Conservation of Feeling

Claim: Across any unified conscious stream, all pleasure and pain accumulate in a ledger that must close to neutral at the death of mind. This asserts a conservation-like principle for subjective experience: the cumulative hedonic balance of any life nets to zero at its end.

Implications: Resources like money or years of life matter only insofar as they affect felt experience. The standard of fairness here is the integral of affect over a lifetime, not the distribution of external goods or opportunities. This shifts focus to suffering and relief as the core variables. Interventions that reduce distress or enable comfort (palliative care, mental health support) matter because they change what is actually tallied.

Signatures: We would expect evidence that life ledgers converge as lives draw to a close. In longitudinal studies, variance in cumulative HCU across individuals should narrow near the end of life. We also look for specific balancing routes predicted elsewhere in the book: dream counterweights and horizon-driven intensity shifts. These are observable manifestations of a conservation-style closure claim.

Fail pattern: Well-measured conscious streams that end with their ledger clearly outside the neutral band despite intact channels for compensation. If lives end definitively positive or negative beyond $\pm K$ HCU under adequate measurement and open relief channels, this assertion is violated.

3.2.2 Universality

Claim: The Law of Fairness applies to all entities that meet the unity criteria for having a conscious stream. No unified conscious stream is exempt from the constraint.

Implications: A universal claim requires comparability. We cannot declare universality unless “1 HCU” is meaningfully linked across cultures, ages, and contexts, and, where applicable, across species. This makes measurement invariance central rather than optional. It also forces conservative inclusion criteria for non-humans: we must specify what evidence of unity and valence is sufficient to say a candidate entity has a ledger at all.

Signatures: The core phenomena (neutral closure, QS-linked signatures, dream counterweights) should replicate across groups once measures are properly linked. Cross-cultural studies should show the same closure behavior when invariance checks pass. If certain non-human mammals qualify as unified streams, we would look for analogous patterns using appropriate proxies.

Fail pattern: Stable group-specific departures from neutrality that persist after careful calibration and stream identification. If a category of genuine unified streams reliably ends outside the neutral band, and we can rule out measurement artifacts, universality fails.

3.2.3 Neutral Death

Claim: At termination of a conscious stream, the lifetime ledger's prespecified confidence (e.g., the 90% interval used for TOST at $\alpha = 0.05$) or credible interval lies entirely within a small preregistered neutral band around zero. In shorthand, every life ends effectively neutral in net felt experience.

Implications: End-of-life measurement becomes central. We are not only interested in how someone feels in their final minutes; we must estimate the integral up to closure. This requires dense monitoring in the final phase, careful adjudication of consciousness so we know when to stop the ledger, and equivalence testing rather than null-hypothesis testing.

Signatures: Ledger dispersion compresses near the end of life. Horizon-linked compensations intensify as closure approaches, conditional on intact compensatory channels. In groups with shrinking horizons, we expect stronger opposite-valence dynamics than in comparable groups not near closure.

Fail pattern: Expanding or persistently wide ledger variance near death, or a substantial fraction of adequately measured streams with final ledgers outside the neutral band. Systematic non-neutral outcomes under intact measurement are direct evidence against this assertion.

3.2.4 Fairness as Design

Claim: Apparent chaos in who suffers and who thrives is compatible with lawful equilibrium at the life level. The balance is not achieved by intention or cosmic purpose but by a constraint on possible trajectories. No life can end permanently imbalanced.

Implications: If the Law holds, it should operate through ordinary channels: rest, repair, social reconnection, symptom relief, and everyday forms of meaning-making. Ethically, this reinforces a practical focus on keeping humane channels open. It does not license

complacency. The constraint, if real, would be realized through distributed processes, not storybook justice.

Signatures: Balancing dynamics appear as low-drama compensations: recovery, relief, reconnection, and other mundane mechanisms rather than miracles. Continuous well-being telemetry should show that compensation typically arrives through ordinary physiological, psychological, and social routes. Dream counterweights, if present, should be symbolic and regulatory rather than prophetic.

Fail pattern: If putative balancing appears only through spectacular external interventions, or if distributed compensation is absent in normal data, then the “constraint-like” premise weakens. If neutrality would require overt cosmic intervention, or if nothing lawful emerges in measurement, this assertion fails.

3.2.5 Resolution of Extreme Suffering

Claim: The existence of suffering, even extreme suffering, is not evidence against life-level fairness. Every pain is an entry in a ledger that, under the Law, must be counterbalanced within that same stream before the death of mind. No stream ends with an unpaid pain-debt.

Implications: This does not trivialize suffering or suggest we can ignore it. Present suffering demands present care. If the Law is true, compassionate interventions are among the normal routes by which balance is restored. The claim is descriptive, not prescriptive, and it never excuses harm or neglect.

Signatures: After sustained negative drift, we expect elevated probability of opposite-signed experiences arriving through plural, humane channels: effective analgesia, reconnection, laughter, intimacy, creative flow, restorative sleep, and related sources of relief. Critically, what matters is relief for the sufferer, not external justice or vengeance.

Fail pattern: Chronic suffering that persists unmitigated until death despite open channels for relief and credible evidence of no compensatory counterweights. If individuals die with large, well-measured surplus suffering that cannot be explained by measurement failure or closed channels, this assertion is falsified. Likewise, if apparent “resolution” is merely habituation to misery rather than genuine positive counterweights, the Law’s mechanism is not supported.

To evaluate these cases properly, lifetime experiences must be measured comprehensively, including small but real sources of comfort and ease: bodily relief, pleasant meals, laughter, intimacy, love, daydreaming, and other restorative episodes.

The Law is falsified if prolonged suffering consistently fails to resolve within the same conscious lifetime under adequate measurement.

3.2.6 Mechanistic Regulation

Claim: There exists a constraint-weighted regulatory system, the Queue System (QS), that biases available thoughts and actions so that trajectories making neutral closure impossible are gradually removed from consideration. This bias strengthens as expected remaining time (horizon) shortens. QS quietly prunes paths that would leave the ledger unbalanced and increases weight on paths that keep closure attainable, especially near the end.

Implications: From the first-person point of view, this does not feel like coercion. It appears as somatic markers: gut feelings, emotional pulls, and intuitive urges that guide behavior. Agency remains intact. People still choose among options, but the salience of options is shaped by QS weights.

From the outside, QS should appear as faint but reliable behavioral and neural patterns. As horizons shrink or a ledger skews, we should see growing tendencies toward restorative behavior: seeking comfort, reconciling, making amends, accepting help, even when such actions are not maximally rewarding by conventional criteria. In valuation and impulse-control circuits, we should see stronger signals when considering actions that worsen imbalance versus actions that relieve it.

When corrective processes are suppressed too long, larger compensations may emerge. In the proposed model, large sustained surpluses in either direction predict stronger opposite-valence correction later. QS preserves agency while tuning choice weights by compensability: how much an action improves the probability of finishing neutral.

Signatures: We test for QS-residuals: traces in behavior or neural data indicating something beyond standard reward-punishment learning. In decision studies, after controlling for established predictors (utility/benefit U, conflict C, arousal, salience), there should remain a significant effect of compensability Φ on choices. Operationally, Φ can be modeled as a function of recent ledger drift and remaining horizon. The core prediction is:

$$A(t) = \beta_0 + \beta_1 U + \beta_2 C + \beta_3 \text{Arousal} + \beta_4 \Phi + \varepsilon$$

If the Law holds as stated, we expect $\beta_4 > 0$: higher compensability biases control-hub activity and policy selection toward repair options.

Additional signatures include horizon \times compensability interactions. As horizons shrink, the probability of choosing repair options should increase relative to baseline. Under

explicit time-short cues, choices may tilt toward actions with long-horizon emotional payoff (reconciliation, relief, acceptance of help) even when short-term costly.

Fail pattern: If careful studies show no trace of these effects, for example $\beta_4 = 0$ after modeling known drivers, no horizon interaction, or all apparent effects fully explained by standard accounts (reward maximization, homeostasis, risk aversion, social norms), then mechanistic regulation fails. If rival models account for the observations without a QS term, adding QS is unnecessary.

Context: Conservation fixes the target (neutral total); Universality the scope (all unified streams); Neutral Death the endpoint; Fairness-as-Design the mode (constraint, not intention); Resolution of Extreme Suffering handles hard cases; and Mechanistic Regulation supplies testable machinery (QS). If any of these assertions systematically fails under robust tests, the Law of Fairness collapses.

3.2.7 Where we go next:

We now have the assertions and their predicted signatures. Section 3.3 draws hard boundaries around the claim, clarifying what the Law does not say, so our tests and discussions are not confused with karma, cosmic justice, or “fair on average.”

3.3 What the Law Does Not Say

A strong claim needs sharp edges. “Fairness” is easy to misread as comfort, justice, or metaphysics. Here are the boundaries. The Law of Fairness, as used here, does not imply any of the following:

- Not a promise of easy lives. The Law does not say that day-to-day life will be gentle, that events will be just, or that suffering will be rare. It speaks only to the lifetime ledger at closure, not to the smoothness of the path getting there. Terrible things can and do happen. The claim is only that, at the end, the total is neutral within preregistered bounds. This is cold comfort, and it does not make the pain of the moment any less real or any less in need of relief. Balance at the end is never an excuse for pain in the middle. Present pain warrants present care.
- Not moral desert or cosmic justice. The Law ties “fairness” strictly to felt experience, not to virtue, vice, or moral merit. It does not guarantee that good people live pleasant lives or that cruel people suffer for their deeds. There is no moral scorekeeping. A neutral ledger does not mean “just deserts.” It is a balance of experiences, not ethical reward and punishment.
- Not “fair on average.” The claim is not that the average across many people is near zero. A population mean could be near zero while individuals end far from neutral. The Law is about each unified conscious stream. If even one well-measured life ends outside the bounds, that is a problem for the Law. This is a life-level claim, not a statistical trend across groups.
- Not teleology or intention. We are not saying “the universe wants balance,” nor that an intentional force guides lives toward fairness. If the Law holds, it functions like a constraint: it shapes which histories occur without any aim or plan. “Law of Fairness” is shorthand for a restriction on allowable trajectories, not a cosmic purpose.
- Not karma or afterlife bookkeeping. No reincarnation, eternal souls, or divine ledgers. The Law is this life, this stream. If a stream ends (death of mind) with uncompensated pain or pleasure outside the preregistered band, the Law fails. There is no hidden repayment. No metaphysical escape hatches.
- Not a license to ignore suffering. Most importantly, “balance at the end” is never an excuse in the middle. We do not condone telling someone in pain, “Don’t worry, it will all balance out later.” Present pain deserves present care, without diminution. If the Law holds, care and relief may be among the channels through which balance is achieved. If the Law is false, the duty to alleviate suffering remains unchanged. Under no interpretation does this hypothesis permit neglect or cruelty. Comfort and dignity override data. We never withhold analgesia,

therapy, or human kindness for the sake of an experiment or theory. The ethos is simple: relief is a systems variable.

- Not a mere tendency toward fairness. Adaptation, opponent-process rebounds, and hedonic-treadmill effects can aid recovery, but tendencies are not laws and they can leave individuals unbalanced. The Law, if true, is stricter. It is not “most people even out.” It is neutral closure for each stream.
- Not a denial of agency. A constraint or a QS mechanism does not mean people are puppets. Individuals still deliberate, err, learn, and choose. The constraint shapes the feasible menu and subtly weights options, but it does not select for you. It is akin to how gravity constrains movement without deciding whether you climb a mountain or stay home. From the inside, this would feel like affective pulls and somatic markers, not external coercion.
- Not equality of goods or outcomes. The Law does not claim equal wealth, equal opportunity, equal health outcomes, or any other material or social equality. Those forms of fairness, important as they are in ethics and policy, operate on different metrics. External circumstances are often grossly unfair. The Law’s claim is confined to the subjective aggregate of each life. It does not require equal incomes, lifespans, achievements, or social justice. A life can be objectively tragic or oppressed and still, in principle, end with a neutral ledger. None of this reduces the urgency of pursuing justice in external dimensions. This hypothesis is not a substitute for social reform.
- Not reduction of meaning to mood. Focusing on felt experience does not imply that meaning, dignity, love, or purpose are “just mood.” They are among the most consequential contributors to the ledger precisely because they are felt. Losing meaning registers as suffering; gaining it registers as fulfillment. Treating feelings as the final currency is not a debasement of value, but an insistence that value be counted in the only coin that matters to the subject: how life feels. The Law speaks to the net total, not to the richness, structure, or story of a life.
- Not omniscient measurement. Declaring the Law does not imply perfect measurement of happiness or suffering. We cannot “read a soul,” and we do not pretend to reduce a life to a single number with certainty. HCl and HCU are composites with uncertainty intervals. Many individual cases may be too noisy or incomplete to classify. The claim is that, where measurement is adequate, neutral closure should appear within preregistered bounds, and that methods can improve over time without relaxing the criterion.
- Not retroactive edits. We will not salvage the Law after the fact by changing who counted as a person, when the stream “really” ended, or what “neutral” should mean. Unity criteria, death-of-mind adjudication, and equivalence bounds ($\pm K$)

are fixed beforehand through preregistration and prior chapters. If results are unfavorable, we take the hit. No moving goalposts. No “No True Scotsman” escapes. If the Law cannot survive fixed definitions, it should not survive.

- Not a denial of inequality or injustice. Structural injustices are real. Oppression, exploitation, and random misfortune happen and cause immense suffering. The Law, if true, would say that these experiences are somehow offset for the individual by the end. It does not justify injustice or suggest it is acceptable. If anything, it implies that balance, if it exists, likely arrives through care and repair rather than through denial. And if the Law fails, the moral imperative to reduce injustice remains as strong as ever. This is a hypothesis, not a moral principle, and never a substitute for compassion or equity.
- Not a happiness-maximization principle. The Law is not a utilitarian rule that life “tries” to maximize happiness or minimize pain. A system can be stable in homeostatic or predictive terms and still permit a life to end unhappy. Standard reinforcement learning and predictive processing optimize local signals; they do not require final cumulative reward to equal zero. Our claim is different: an endpoint neutrality constraint, not a drive toward maximal happiness.
- Not panpsychism or AI mysticism. We restrict the claim to streams that meet explicit scientific criteria for consciousness. We are not implying that the universe as a whole is conscious or that every AI system has feelings. “Unified conscious stream” is operationally defined (via the Unity Index criteria). If something does not meet those criteria, it is not in scope. We do not speculate about vague cases such as “the consciousness of the universe,” nor do we assume subjective experience in systems without evidence.

3.3.1 Where we go next:

With misreadings fenced off, Section 3.4 sets the edge cases: when ledgers pause, split, or merge, and how identity is adjudicated. Those boundary conditions keep the tests honest when real lives do not fit cleanly.

3.4 Boundary Conditions

A law is only as credible as its behavior at the edges: corner cases where definitions strain and failure would be easiest to hide. As in physics, we test laws under extreme conditions. Here we stipulate the boundary conditions for the Law of Fairness: who and what it applies to, when the ledger runs versus pauses, how fragmentation and fusion are handled, what counts as the end of a stream, and how we treat extremely short lives or non-human candidates. Crucially, we declare these boundaries before examining outcome data. If the Law fails under fixed rules, we want that failure to be unmistakable, not blurred by elastic definitions.

3.4.1 The bearer criterion: Unity by Access

Rule: The ledger accrues only when a unified conscious stream is present. Operationally, we require a Unity Index to exceed a threshold θ to consider consciousness “on” and unified. The Unity Index has three pillars:

Access: Is information globally available to the agent for guiding behavior or report?
Integration: Are sensory, cognitive, and affective signals bound into a coherent state rather than disjoint fragments? Control coherence: Do actions and decisions indicate a single integrated agent rather than multiple competing sub-agents?

We set a cutoff value θ on the Unity Index. The threshold may differ by context (for example, healthy adults versus infants or animals, or specific anesthetic regimes), but it is prespecified for each context. In practice, θ is anchored to clinical or normative criteria, such as minimal reliable signs of awareness on established scales. The key point is that θ is fixed in advance to avoid post hoc adjustments.

Non-verbal ledgers: Where self-report is impossible (neonates, advanced illness), we estimate $F(t)$ from composite indicators such as facial action or EMG, heart-rate variability, oxygen saturation, pain-indexed EEG features, acoustic distress markers, and caregiver or clinician ratings. These enter as first differences into $HCI\Delta$ with reliability weights and measurement-invariance checks. This keeps $L(t)$ auditable for non-verbal lives and prevents exclusion of the very cases critics raise.

If Unity Index $\geq \theta$, ledger accrual is active. If Unity Index $< \theta$, the stream is considered absent (or not unified enough) and ledger accrual pauses. In essence, no ledger entries accumulate when consciousness is not sufficiently unified.

Fail pattern: If the Unity Index is unreliable, the framework loses a stable subject. Examples include frequent classification reversals without real state change, poor inter-rater reliability for the same state, or systematic non-generalization across populations.

In that scenario, we cannot consistently define whose ledger we are summing. This is a precondition failure more than a direct falsification of neutrality, but it is critical: if we cannot identify the bearer of experience, we cannot evaluate the Law. The appropriate response would be to revise the unity measure or acknowledge that the Law cannot be tested reliably until the bearer criterion improves.

3.4.2 Pauses: Sleep, anesthesia, coma

Rule: During any state where the conscious stream truly disappears, such as deep dreamless NREM sleep, sufficiently deep general anesthesia, or true coma, the Unity Index $< \theta$ and ledger accrual is paused. No HCU entries accumulate during these periods because there is no unified experience to have valence. Time spent utterly unconscious does not count toward the ledger; the account is on hold.

Exceptions: If the stream resumes in some form during these states, we count it. REM sleep and other phases that include organized dreaming are special cases. If evidence suggests structured dream experience (Unity Index above a prespecified dream threshold), a dream stream is active and ledger accrual resumes. Dream affect can contribute positive or negative HCU and may participate in counterweight signatures (Chapter 10). The guiding principle is simple: if there is evidence of unified conscious experience, the ledger runs; if not, it does not.

Fail pattern: If research shows that periods we treat as “paused” actually contain structured valenced experience that later influences the person as if it were felt, then the pause rule is wrong and must be refined. A related implementation risk is threshold error: if our θ is set too high, we may misclassify low-consciousness states (e.g., minimally conscious) as pauses and fail to count suffering. To mitigate this, we err conservatively: in borderline cases, assume some consciousness and count, rather than risk erasing experience. If the pause assumption proves systematically false, the accounting framework must be revised before the Law can be fairly tested.

3.4.3 Fragmentation: Split-brain, dissociation, delirium

Rule: When a single organism’s consciousness splits into two or more concurrently active streams that do not share unified awareness, we treat this as branching of ledgers. Examples include rare split-brain cases, severe dissociative cases with strong amnestic separation, or certain delirium states that plausibly reflect alternating agencies. In these scenarios:

- Concurrent branches: If two or more streams run in parallel in one body, and each has Unity Index $\geq \theta$, each stream accrues its own ledger while separate.

- Alternating branches: If only one stream is active at a time, but there is true amnestic separation between streams, the active stream accrues while inactive streams pause. Experiences do not mix if awareness is not shared.
- Reconciliation: If branches later reunify into one stream (integration in DID therapy, functional reintegration after injury, rejoining after partition), the ledgers fuse by summation. We sum HCU across branches and propagate uncertainty appropriately. We do not retroactively discount one branch's suffering because another branch "did not feel it." If it was felt by a stream, it counts.

Fail pattern: Stable plural consciousness in one organism that never reconciles is a hard edge case. Under our rule, it becomes two ledgers, and the Law is evaluated per stream: each stream must still close neutral at its own death of mind. If one branch ends clearly non-neutral, the Law fails for that branch. If such cases exist widely, they stress the formulation and the endpoint definition (one mind could cease while another continues in the same organism). If those realities routinely break neutral closure, the Law would fail or require a serious revision that makes multi-ledger closure explicit rather than ad hoc.

3.4.4 Fusion: Recovery of unity

Rule: When a previously fragmented consciousness regains unity, we treat it as one stream again going forward. Examples include severe dissociative cases that integrate into a single self (through therapy or spontaneously) and brain-injury cases where divided functions reintegrate. When the Unity Index rises above θ and remains above θ for a sustained period (we may require a stabilization window, for example 72 hours of continuous unity, to reduce false positives), we declare that the streams have fused.

At fusion, we merge the ledgers. The unified stream's ledger becomes the sum of all branch ledgers accrued up to that point, plus new entries going forward. Uncertainty from separate phases is combined under the prespecified dependence structure: if branches were independent over disjoint intervals, uncertainty can be combined in quadrature; if branches overlapped or interacted, we incorporate covariance where estimable.

Practically, for a patient who "wakes up" into unity after fragmentation, we sum the HCU accrued by each fragment during its active periods. That sum becomes the baseline ledger for the unified person. We do not retroactively discard any branch ("that alter's suffering doesn't count now"). If it was experienced by a qualifying stream, it counts. This ensures that nothing felt is erased by bookkeeping.

Fail pattern: The accounting assumption is summation. If recovery of unity reliably negates or erases the impact of what occurred during fragmentation, our summation rule may misrepresent what “the life” ultimately contains. For example, suppose a person had two streams A and B, and after integration the unified person has no access to B’s memories and B’s past leaves no psychological or physiological trace. One might argue that B’s ledger should not fully count if B’s experiences are functionally irretrievable and causally inert in the unified person’s later life.

We deliberately do not drop B’s ledger, because it was felt at the time by a qualifying stream. However, if integrated individuals systematically show zero trace of one branch’s past, it raises a conceptual challenge for testability: a forgotten pain that leaves no mark still counts by our definition, but may be empirically invisible. A stark failure mode for the measurement program is this: someone splits, one branch suffers greatly, then fuses, and the unified person shows no detectable residue of that suffering. In that case, the ledger approach faces an empirical wall. Either we include the hidden ledger by principle (risking the appearance of special pleading) or we exclude it (risking an ad hoc “escape hatch”). Persistent, systematic “loss of ledger” upon fusion would therefore be a conceptual failure for our ability to test the Law rigorously. It would imply there are experiences that contribute to the ledger yet have zero measurable effect on the final observable state, complicating strict neutrality adjudication.

3.4.5 Tiny ledgers: Early loss and minimal streams

Rule: For very short lives, for example a miscarriage that briefly attains consciousness, a stillbirth, or a neonatal death after hours or days, the normal machinery of the Law is harder to apply. The Unity Index may never clearly rise, or may do so only transiently. We therefore define special handling:

- No-unity cases: If a being never achieves $\text{Unity Index} \geq \theta$ (for example, no neurological basis for awareness such that reflexes occur but unified experience is implausible), then there is no ledger to speak of. By definition, there is no conscious accrual to evaluate. We exclude such cases from analysis, not because they “do not count morally,” but because the Law is about conscious streams. No stream, no ledger, no neutrality test.
- Minimal unity cases: If a conscious stream exists but only briefly (minutes, hours, days), empirical adjudication of neutrality is severely underpowered and uncertainty dominates. The ledger can accrue, but classification will usually be Inconclusive unless something extreme occurred. For these cases, we may apply a prespecified scaling rule that broadens the effective equivalence band as a function of total conscious duration T. The intent is not to “save” the Law, but to

reflect the reality that with tiny T and sparse data, small absolute imbalances are not meaningfully testable. The scaling rule, including any cutoffs and multipliers, must be preregistered. The shorter the conscious life, the more cautious the neutrality verdict.

Fail pattern: If short-lived streams show a large, systematic skew beyond even the broadened bounds, reliably measured, that would refute the Law for that category or force a revision of θ and the scaling policy. Although robust measurement here is ethically and practically difficult, the logic is clear: if newborns who die shortly after birth consistently exhibit distress with no compensatory positives, and this pattern can be established under humane measurement, those cases are failures of the Law in that domain.

A methodological failure is also possible: if broadening K becomes so permissive that it renders tiny-ledger cases unfalsifiable, that is bad science. Any scaling must be principled, duration-linked, and specified in advance. If tiny-ledger patterns emerge that remain non-neutral beyond uncertainty bounds, we must accept either that the Law fails at the margins or that such cases are explicitly out of scope.

3.4.6 Long plateaus and vegetative states

Rule: Some individuals enter prolonged states of disordered or minimal consciousness (e.g., Unresponsive Wakefulness Syndrome, formerly “vegetative state,” or Minimally Conscious State). We handle these by continuous assessment rather than assuming fully “on” or fully “off.” The Unity Index is applied using multimodal evidence: behavioral scales (e.g., CRS-R), EEG/ERP responses, active paradigms where feasible, and neuroimaging markers when ethically appropriate. Ledger accrual is active only during periods with evidence meeting Unity Index $\geq \theta$. If Unity Index remains below θ , the ledger is paused even if the body appears awake. If signs of consciousness emerge intermittently, accrual resumes during those windows.

Fail pattern: The central risk is covert consciousness. If our methods fail to detect awareness, we may incorrectly treat the ledger as paused when the person is experiencing distress. That is an implementation failure that can mask non-neutral outcomes. We mitigate this by using the most sensitive methods available and erring conservatively: in borderline states, assume some consciousness and keep the ledger open rather than closed.

A direct challenge to the Law occurs if a person remains minimally conscious for years with strong evidence of suffering and then dies without compensation despite available

time and humane relief opportunities. Our boundary rules do not create a loophole for that case. If confirmed under adequate measurement, it is a counterexample.

3.4.7 Neurodegeneration and identity drift

Rule: In progressive neurological disorders (e.g., Alzheimer's disease, frontotemporal dementia), unity may gradually degrade. As long as Unity Index $\geq \theta$, the ledger accrues. If Unity Index drops below θ and remains below in a sustained way, we treat the ledger as closed at the point of irreversible loss of unity. If there are later reliable crossings above θ (transient lucidity or fluctuating awareness), accrual resumes during those intervals.

Because these patients can hover near threshold, we may impose a stabilization window to avoid rapid toggling. For example: if Unity Index remains below θ for a continuous prespecified interval (e.g., one week), we classify the stream as ended for ledger purposes unless strong evidence of return emerges later. This prevents impractical and ethically fraught “daily opening and closing” of a ledger in late-stage decline.

Compassion clause in analysis: We explicitly report how much ledger time accrued before prolonged sub-threshold periods, and we treat the tail with heightened transparency about uncertainty. We do not want to over-interpret “neutral at death” if the verdict relies on long periods where experience is ambiguous or poorly measurable. Where appropriate, extended sub-threshold periods may be treated as censored or analyzed with sensitivity checks. The Law should hold even under that uncertainty; we simply refuse to overclaim.

Fail pattern: If many cases follow the same pattern—severe suffering during the final clearly conscious phase followed by loss of unity and then death—the Law faces a hard test. We do not count mere non-experience as compensatory pleasure; absence of suffering is not addition of happiness. Under our criteria, if the last conscious interval ends strongly negative and then the stream ceases, the ledger ends non-neutral unless genuine counterweights occurred before closure. If such trajectories are common in neurodegeneration, the Law would fail frequently in this domain unless comfort measures reliably produce measurable relief while unity remains present.

3.4.8 Resuscitation, NDEs, and oscillatory returns

Rule: If clinical death occurs (e.g., cardiac arrest) but the person is resuscitated and conscious signs return meeting Unity Index criteria, we treat this as the same stream continuing. The ledger did not close at cardiac arrest; it remains one continuous account with an interruption. The endpoint T is the later time when unity finally ceases.

Near-death experiences (NDEs): Reported experiences during apparent clinical death are counted only if there is reason to believe consciousness was active then. Often, vivid NDE imagery may be generated during the transition into shutdown or during recovery rather than during a flatline interval. We lean conservative: absent strong evidence that Unity Index met threshold during the interval in question, we do not assume sustained conscious experience in the middle of complete physiological cessation. If evidence indicates conscious experience occurred, we count it with its valence like any other period.

Closure moment: If resuscitation fails or is not attempted, death of mind T is the last moment unity was above 0. Closure may precede cardiac arrest (if brain shutdown leads), coincide with it, or follow briefly (if a final surge sustains unity for seconds). The operational rule is consistent: closure is keyed to unity, not to legal time.

Fail pattern: Most risks here are interpretive rather than conceptual. Repeated oscillations in and out of consciousness complicate measurement but do not change the rule: if unity is present, the ledger runs; if not, it pauses; closure occurs at the last sustained loss of unity. NDE content is counted only when unity is plausibly present. If NDEs systematically deliver strong positive or negative affect near potential closure, they become part of the ledger and can, in principle, contribute to or challenge neutrality. The Law stands or falls on the final ledger at true closure, not on the presence of an NDE narrative.

3.4.9 Multiperson cases: Pregnancy, transplant, conjoined twins

Rule: The ledger attaches to a conscious stream, not to a particular body part or biological substrate. This clarifies cases where biology overlaps:

- Fetus/Newborn and mother: Once a fetus reaches a stage where consciousness is plausible and unity criteria could be met, it would count as a separate stream. The mother's ledger is distinct. Pregnancy does not merge ledgers. If the fetus or newborn dies, that tiny ledger is its own case (see 3.4.5). The mother's feelings about the event are entries on the mother's ledger.
- Organ transplant: No ledger information transfers with an organ. Any psychological changes in the recipient are entries on the recipient's ledger, because the recipient is the one who experiences them. The donor's stream, if ended, closed at that time. We reject "tissue carries consciousness" interpretations; the stream is tied to the integrative system that supports experience, not to individual organs.

- Conjoined twins: If there are two brains yielding two conscious minds, there are two streams and two ledgers, even if organs are shared. Each must satisfy or fail the Law independently. If one twin dies, the other's ledger continues, except for the survivor's emotional experience of loss.

These scenarios mainly clarify application rather than create unique failure modes. Failures are ordinary: if a qualifying stream ends non-neutral, that stream is a counterexample regardless of shared biology.

3.4.10 Non-human animals

Rule: The Law may extend to non-human creatures only if they meet rigorous, species-appropriate criteria for a unified conscious stream. We require evidence of unified access, goal-directed control coherence, and valence. If an animal plausibly qualifies (e.g., certain mammals), we include it only if we can construct a calibrated HCl using ethological proxies and physiological markers anchored to known positive and negative stimuli, with explicit uncertainty.

Because animal measurement is difficult, error bars will be large, but the principle remains: if it truly has a stream, the Law claims it should close neutral within prespecified bounds appropriate to that measurement regime.

Fail pattern: If a species with strong evidence of unity shows systematic end-of-life ledgers outside the neutral band after careful calibration and comparability checks, that challenges universality. If results conflict and we suspect the cause is measurement non-equivalence, exclusion may be warranted on scope grounds, but that itself limits universality (as addressed in Section 3.2.2). A robust pattern of non-neutral outcomes in a clearly conscious species would be a serious challenge.

3.4.11 Cultural and linguistic variance

Rule: When comparing ledgers across cultures or languages, or aggregating data globally, we require measurement invariance. At minimum, metric invariance must hold so that a one-unit change in HCl represents comparable affective change across groups. Where scalar invariance fails, we either apply prespecified alignment corrections or propagate the resulting uncertainty rather than pooling incompatible scales. If invariance cannot be established even after careful design and analysis, we do not pool those groups. We restrict conclusions to within-group tests.

Fail pattern: If stable, unresolvable non-equivalence persists, we must suspend cross-group conclusions. That does not automatically refute the phenomenon, but it prevents an empirical claim of universality “for all humans” on a single scale. If one group shows

neutrality and another does not, and we cannot determine whether this is real or a measurement artifact, the correct response is either improved measurement or an explicit limitation on scope.

3.4.12 Catastrophe and forced endpoints

Rule: The Law does not guarantee a long life or advance notice. Sudden deaths (accident, acute medical event, violence) are valid closure points T. The claim is that among life trajectories that actually occur, even abrupt closures end with ledgers inside the preregistered neutrality band. The Law does not promise time to “tie up loose ends” or a final epiphany. It asserts neutrality at T.

Implications: An unexpectedly brief life can still satisfy neutrality, either because little imbalance accrued or because compensations already occurred earlier. Lack of warning is not a loophole. It removes one route to neutrality (late-horizon dynamics), but the Law stands or falls on the ledger at closure.

Horizon clause (mechanism vs. law): Short horizons amplify compensatory pressures when they exist; catastrophes eliminate that window. The absence of a final surge is not evidence against the Law. Evidence against the Law is a non-neutral ledger at T.

Mass catastrophes: Large, abrupt events (bombings, natural disasters) remain testable. Within this framework, three coherent readings exist:

1. Constraint view: only trajectories already lawful to close (already neutral) terminate at the instant; others do not.
2. Selection view: instantaneous deaths preferentially occur among those already near neutrality, while those far from neutrality persist longer.
3. Measurement view: if adequate measurement shows instantaneous closures outside the band, the Law fails.

None of these readings justifies tragedy. The Law is descriptive, not moral.

Test. Treat each catastrophic death as an ordinary terminal test: same equivalence bounds, same unity criteria, same error checks. Examine distributions. If sudden closures cluster within bounds under adequate measurement, the Law survives this domain. If they do not, it fails.

Fail pattern.

- Single-case failure: a sudden death with a well-established unified stream and adequate lifetime measurement ends outside the neutrality band.

- Pattern failure: sudden deaths as a class systematically end outside the band; “no warning” does not rescue this.
- Model redundancy: rival accounts reproduce observed patterns without a neutrality constraint, eliminating explanatory necessity.

Bottom line. No warning confers no free pass. We check the ledger at closure. If many abrupt deaths end with uncompensated imbalance under adequate measurement, those cases are decisive evidence against the Law.

3.4.13 QS scope and limits at boundaries

Rule: QS is posited to bias within what a person can do or feel. It has limits. Near hard constraints—extremely short time, severe impairment, unavailable resources—QS cannot create options ex nihilo. It can only reweight among what exists.

Concretely, QS adjusts choice weights; it does not conjure new channels. If biology or environment blocks compensation channels (no access to analgesia, severe physiological constraints, etc.), QS cannot override that hard limit. Near end of life or in constrained states, QS may strongly weight whatever compensatory paths remain, but it cannot make an impossible path possible. The Law, if true, remains a universal constraint; what changes is how the constraint could be satisfied and how well we can detect it.

Fail pattern: If compensable options exist and are routinely ignored without any detectable tilt toward relief as closure nears, QS is undermined. For example, if near death many people decline available pain relief and show no alternative compensatory dynamics, and this is systematic rather than idiosyncratic, it weakens the QS claim. Likewise, if horizons shrink and behavior does not show any reliable shift toward repair options even when available, mechanistic regulation loses support. The Law might still be true in principle, but the proposed mechanism would be in trouble.

3.4.14 Equivalence bands and tiny/huge T

Rule: The neutrality band ($\pm K$ HCU) is set a priori and reflects what counts as negligible lifetime imbalance. K is not adjusted post hoc. However, edge guidelines apply:

- For extremely short lives: As discussed in 3.4.5, we may prespecify a duration-linked scaling that broadens the effective band when total conscious duration is tiny and uncertainty dominates. This must be declared in advance.
- Very long lives: For long-lived streams (hypothetical extended lifespans), we generally keep K fixed in absolute HCU units because K is anchored to meaningful

effect sizes. “Material imbalance” should mean the same subjective impact regardless of life length.

We also run preregistered sensitivity analyses around K (e.g., $\pm K/2$ and $\pm 2K$) as secondary checks. The primary K remains the basis for conclusions.

Fail pattern: If neutrality conclusions flip under small, reasonable variations of K, the result is fragile. That does not automatically falsify the Law, but it undermines confidence and invites the concern that K is arbitrary. We counter this by anchoring K to interpretable effect sizes and preregistering it. A methodological failure would be post hoc band selection to make results pass. We do not do that. If the Law “passes” only for conveniently large K and fails for any stricter plausible definition of neutral, that is, in practice, a weak or partial success at best.

3.4.15 Missing data, silent stretches, and burden

Rule: Real longitudinal measurement contains missingness: device outages, nonresponse, opt-outs, and end-of-life constraints. We treat sustained missingness as part of the process rather than ignoring it. If missingness correlates with distress (e.g., participants drop monitoring when they feel bad), we model it explicitly (pattern-mixture or selection models) and widen uncertainty accordingly. During silent stretches, we do not “fill in” the ledger with wishful interpolation. We propagate uncertainty.

If uncertainty in L(T) remains too large to classify, we declare the case Inconclusive. This is an integrity rule. It is better to say “we cannot tell for this person” than to force a verdict.

We also track measurement burden. We will not compromise comfort to reduce missingness. If missingness occurs because monitoring was paused for dignity and relief, that is acceptable. We then report the widened uncertainty as part of the scientific record.

Fail pattern: Missing data does not itself falsify the Law, but it can prevent a fair test. If a large fraction of cases end inconclusive due to missingness, the empirical program may fail to adjudicate the claim in practice. This is a practical failure of testability, not proof that the Law is true or false. It motivates improved, lower-burden measurement and transparent reporting of indeterminacy.

3.4.16 Multi-agent coupling and shared channels

Rule: Lives are not isolated. Many compensation channels are social and many resources are shared (caregivers, analgesics, safe environments). The Law is claimed per stream, but its realization would occur in a coupled world. If multiple people need scarce resources to achieve relief, they cannot all access the same channel simultaneously. The

Law does not imply that every life is balanced at every moment; it claims neutrality by each stream's end.

We model this as shared channel capacity. Let $R(t)$ be a vector of shared resources with capacities $C(t)$. The admissibility of an action u for person i depends on these resources and on the policies of others. Lives are coupled through competition and cooperation for balancing channels. When shared channels saturate, we predict that menus shift toward available, less-contested compensations (sleep, peer support, self-soothing) and that scarce resources are allocated with horizon sensitivity: those nearest closure may receive priority in practice, as a kind of emergent shadow price on relief opportunities.

This is complex, and Part V returns to social coupling in depth. The key point here is definitional: the Law does not conjure infinite resources. If it holds, it must hold in a world with constraints, where compensation routes can be delayed, substituted, or redistributed over time.

Fail pattern: If systemic deprivation produces many non-neutral endings (famine, medical apartheid, chronic denial of relief), that is a large-scale failure of the Law. Competition is not an excuse. A life ending non-neutral is a failure, period. If the Law only appears to hold under conditions of adequate social cooperation and resource access, then it is more conditional than stated or false in universal form. Part V examines whether universal neutrality is compatible with real-world scarcity at scale.

3.4.17 Where we go next:

Now that the edges are explicit, Section 3.5 pins down the endpoint. We define “death of mind” operationally and ethically so that closure, and any neutrality verdict, can be audited without guesswork.

3.5 The Death of Mind

“Death of mind” is our term for the endpoint of conscious experience: the last moment at which the conscious stream exists. All equivalence testing, and the Law’s ultimate judgment, is evaluated at that moment, not at cardiac death and not at any convenient timestamp. Getting the endpoint right is the difference between a testable law and a poetic idea. Here we clarify what “death of mind” means, why the endpoint must be experiential, how we adjudicate it in practice, and what signatures we expect as closure nears.

3.5.1 What “death of mind” means (plain statement)

Death of mind is the final instant at which the Unity Index remains at or above threshold θ . Immediately after that, the stream is gone. There is no subject having experiences.

This is not necessarily the last heartbeat or the legal declaration of death. It is tied to the cessation of conscious experience itself. In dying patients, death of mind may occur minutes or hours before cardiac arrest if the brain irreversibly loses unified consciousness first. Conversely, there are scenarios where a body is sustained after the mind has unequivocally ended.

To spell it out: it is not defined by heart or lung function per se, and not even strictly by whole-brain death labels if any unified experience lingers. It is the end of anything it is like to be that person.

Methodologically, once death of mind is reached and confirmed with a stabilization window (to ensure no return of unity), we close the ledger there. Clinically, that means we require sustained markers consistent with no consciousness before calling closure. The exact criteria are prespecified by setting, but the logic is constant: no further bodily happenings count because there is no one home to experience them. If the heart stops at 10:00 but unity was absent since 9:55, death of mind is 9:55 and the ledger closes then, even if an official pronouncement comes later.

We make a few exclusions to avoid confusion:

- It is not cardiac death by itself, because the mind may still have activity or may be revivable.
- It is not legal death if there is reason to believe conscious activity persisted past that declaration.
- It is not the end of bodily processes, since cells can metabolize and reflexes can fire without experience.
- It is specifically the end of subjective experience for that stream.

- Practically, we rely on a combination of neurological criteria (unresponsive, flat EEG or brain signals consistent with no consciousness, etc.) to mark death of mind. In an experimental sense, that is the point at which we perform our final ledger evaluation.

3.5.2 Why the endpoint must be experiential

We insist on an experiential endpoint because the Law is about felt experience. If a bearer is still experiencing, however faintly, the ledger is still open. If experience has truly ceased, adding more time or biological activity adds nothing to the ledger.

Using proxy endpoints (for example, “heart stopped for five minutes,” “declared dead,” or an arbitrary calendar age) creates predictable errors:

- Closing too early: Declaring a stream ended when experience still persisted (for example, covert awareness or residual integration) truncates the ledger and omits real entries.
- Closing too late: Counting post-consciousness noise (brainstem firing, autonomic settling) as though it were felt experience corrupts the ledger and can manufacture false “balance.”

So we ground the endpoint in the only thing the Law cares about: the presence or absence of consciousness. This also matches our ethics. We do not treat someone as “done” until we have strong evidence no further experience is possible.

3.5.3 Clinical adjudication in practice

Declaring death of mind is difficult, so we specify standardized protocols by context. All protocols are:

- Prespecified: Criteria and thresholds are fixed in advance.
- Clinician-led: Determinations are made by qualified medical professionals, not by the research team.
- Non-interfering: Observation is embedded in care and never competes with comfort or treatment.

Different settings require different tools:

- Hospice and palliative care: We rely on bedside assessments for responsiveness and awareness, supported by available monitoring. We document sustained absence of command-following and other indicators of awareness and, where available, neurological markers consistent with loss of unity. We declare T when loss of awareness persists beyond a stabilization period specified in advance.

After that point, further bodily activity does not count for the ledger. Sedation complicates adjudication, but we do not withdraw sedation to “test” consciousness. If sedation produces sustained loss of unity with no return, it can mark death of mind for analytic purposes, with explicit documentation.

- Operating room and ICU (anesthesia, coma, withdrawal of support): We use clinical monitoring and, where feasible, EEG-based indices and evoked-response paradigms to detect residual processing. Closure is called only when multiple indicators converge that unity has been lost and will not return. If any plausible sign of awareness remains, the ledger stays open. In planned withdrawal scenarios, unity typically ceases before cardiac arrest; we locate closure at the last evidence of awareness, not at the moment the heart stops.
- Out-of-hospital sudden death: This is noisier because continuous monitoring is rare. We reconstruct the timeline from witness reports, available device records (for example, wearable heart-rate traces), emergency-response documentation, and the failure or success of resuscitation. We remain conservative: we do not attribute post-collapse experience without evidence, and we explicitly record uncertainty in T when it cannot be resolved. The goal is not perfect certainty; it is an auditable, prespecified rule applied consistently.

Across settings, the principle is the same: capture the last plausible moment of experience and avoid both premature closure and post-consciousness overcounting. When in doubt, we err toward keeping the ledger open slightly longer rather than truncating a possible final experience.

3.5.4 Edge contexts and how we call them

Certain scenarios deserve explicit rules:

- REM-like activity just before death: Some dying brains show late bursts of organized activity. If converging evidence suggests structured mentation (dream-like unity above a prespecified “dream threshold”), we count those minutes. Closure is after that final organized experience ends. We do not cut off simply because a patient appears unresponsive. If the best evidence suggests a last flicker of experience, we include it.
- Minimally conscious states and covert awareness: In grey-zone disorders of consciousness, we avoid declaring death of mind until evidence of awareness is absent across modalities. Behavioral unresponsiveness is not enough if neural markers suggest “someone home.” If covert signs persist, the ledger remains open. Closure requires sustained disappearance of such signs under a prespecified protocol.

- Near-death experiences (NDEs): If an NDE is reported after recovery, it is counted only to the extent it plausibly corresponds to conscious content during the episode. Often, vivid experience may occur during transitions into shutdown or during restart rather than during complete physiological cessation. We do not assume experience during a flatline interval without evidence. In any case, an NDE followed by recovery is not closure; it is an interruption. Closure occurs only when unity is lost and not regained.
- Organ donation after circulatory death (DCD): We follow clinical standards for irreversible loss of consciousness and do not count anything after death of mind. The point of closure is the irretrievable loss of unity, typically well before any surgical intervention. We note explicitly: what happens to the body after closure is not part of the ledger, because there is no mind to experience it.

In all edge contexts, the goal is disciplined inclusion: count what can plausibly be experienced and exclude what cannot, using rules fixed in advance.

3.5.5 Horizon scaling near T (what to expect if the Law holds)

If the Law is true, the period approaching death of mind is where some of its strongest signatures should be detectable. The system is near its boundary condition. We therefore predict the following near T:

- Increased probability of opposite-valence experiences, given intact channels: If humane channels are open (analgesia available, supportive contact possible, sleep permitted, spiritual or narrative closure available), we expect an elevated probability of compensatory experiences in the direction needed for neutrality. For a strongly negative ledger, this could look like relief, reconciliation, calm, or lucid settling that occurs more often than chance under comparable conditions. This is not a promise of a beautiful ending. It is a statistical claim about compensatory dynamics when compensatory channels are available. If channels are closed by neglect or constraint, we do not posit miracles. In such cases, the Law faces a harder test.
- Compression of ledger variance across individuals: In longitudinal cohorts, ledger estimates should converge as each person approaches closure. Practically, the cross-sectional variance of $\bar{L}(t)$ should decline as $t \rightarrow T$, relative to earlier phases. If the spread increases or stays flat under adequate measurement, that is evidence against a convergence constraint.
- For those unable to communicate, rely on converging physiology and neural markers: Many dying patients cannot self-report. We therefore look to autonomic and neural correlates consistent with relief or settling, always interpreted

cautiously and never used to override comfort. Under humane care, we expect distress markers to abate rather than escalate into chaotic end-stage strain. These are testable signatures, not moral expectations.

- No evidence of mounting imbalance right at the end under good conditions: In well-managed end-of-life settings, we should not see systematic trajectories that remain strongly divergent up to closure with no convergence dynamics at all. Repeated absence of convergence where channels are open is evidence against the Law.

We state this plainly: if these patterns fail to appear in adequately powered, well-measured end-of-life cohorts under humane conditions, that counts against the Law.

3.5.6 What does not count as experience

To avoid misclassification at end of life, we explicitly exclude phenomena that can occur around death without implying conscious experience:

- Isolated spinal or brainstem reflexes: Movements such as the Lazarus sign, decerebrate posturing, knee-jerk reflexes, agonal gasps, or reflexive vocalizations can occur without cortical integration. They may look dramatic. They do not, by themselves, imply felt experience. Unless Unity Index criteria indicate cortical access and integration, these events do not reopen the ledger.
- EEG artifacts or non-conscious rhythms: Patterns such as burst suppression or slow-wave activity can reflect deep anesthesia or severe dysfunction without awareness. A burst on EEG is not, by itself, a “thought.” Without markers of access and integration, such rhythms are treated as non-conscious. The ledger remains paused.
- Autonomic storms without evidence of awareness: Sympathetic surges can occur due to disinhibited subcortical circuits. Ethically we treat the body for comfort regardless, but analytically we do not count these storms as experienced suffering unless there is evidence the person is conscious.

These exclusions matter because misattributing reflex physiology to experience would distort the ledger and could generate false failures or false “balances.” Once death of mind is reached, bodily events after that point do not affect the ledger.

3.5.7 Uncertainty at closure (how we report it)

Because determining death of mind and the final ledger is delicate, we report uncertainty transparently. For each case, we compile a closure report including:

- Unity Index trajectory around the end, with the final crossing below θ and the stabilization window documented.
- Final $L(T)$ (and $\bar{L}(T)$ where reported) with a 95 percent confidence or credible interval and the resulting classification (Neutral / Non-neutral / Inconclusive).
- Data-quality indicators near the end: missingness, device uptime, sedation status, and any anomalies that increase uncertainty.
- Context such as expected versus sudden death and whether key relief channels were available (analgesia, support, rest). This context does not change the neutrality rule, but it is essential for interpretation and pooled analyses.

If uncertainty is too large to classify under the preregistered band (for example, the interval overlaps both inside and outside $\pm K$), we mark the case Inconclusive rather than force a verdict. This is an integrity rule. Ambiguous cases are documented as ambiguous.

This transparency also prevents selective reporting. Neutral cases should be clearly neutral under the preregistered criterion. Non-neutral cases should be clearly non-neutral. Inconclusive cases should be labeled as such, not squeezed into a narrative.

3.5.8 Anticipated objections (and our replies)

We address common objections to the death-of-mind endpoint:

- “Isn’t this metaphysical?” No. “Death of mind” is operationally defined by brain and behavioral markers of unity. No soul claims, no supernatural events. We mark the empirical end of an information-integrating process by prespecified criteria.
- “What about sedation masking experience?” Sedation complicates inference but not the ethical rule. We never reduce sedation to “check” consciousness. If unity markers indicate ongoing awareness under sedation, the ledger remains open and we count what can be experienced (including relief). If sedation produces sustained loss of unity below θ with no return, the ledger pauses and may close if unity never returns. We treat the person’s comfort as primary and treat unity as the gate for ledger accrual.
- “Could post-mortem brain surges count as conscious?” We count them only if they meet unity criteria. Some late surges may be disorganized shutdown dynamics. If evidence ever shows a surge corresponds to structured experience, those seconds are part of the final ledger. If not, they are excluded. We acknowledge the uncertainty and apply the same rule: unity, not electricity alone.

We keep the scope scientific: count what can plausibly be experienced, exclude what cannot, and resist metaphysical drift.

3.5.9 Fail patterns specific to end-of-life

Observable patterns at end of life that count against the Law include:

- Stable, well-adjudicated closures with terminal non-neutrality outside the band:
If we can confidently locate T and a nontrivial fraction of cases have final ledger intervals clearly outside $\pm K$, those are direct violations.
- No variance compression near closure in longitudinal cohorts: If ledger outcomes do not converge as closure approaches, or systematically diverge, the predicted attractor at neutrality is not manifesting.
- Null relationship between open relief channels and opposite-signed HCU near T:
If availability of compensatory channels shows no measurable association with predicted end-stage counterweights where those channels are intact, the mechanism-level predictions weaken.
- Systematic miscalls of T discovered after the fact: If we repeatedly declare closure and later obtain credible evidence of continued awareness beyond the declared endpoint, the adjudication protocol is flawed. One-off errors can happen. Systematic errors undermine the entire empirical program and could artificially inflate apparent neutrality.

If these patterns appear consistently under adequate power and methods, the Law is on shaky ground or falsified.

3.5.10 Why this section is here

A law without a clear endpoint invites loopholes: “maybe balance comes later,” “maybe it happens somewhere unseen.” By fixing fairness to the observable end of experience, we make the claim auditable. Anyone can examine the ledger at the death of mind and see whether it falls within the preregistered band or not.

If life-level neutrality is real, it should show up at that line, not as a comforting vibe in the middle of life and not as an average across people, but as a strict closure condition for each stream. If it is not real, the failure should be unambiguous: ledgers ending outside the band, with confidence, and no metaphysical place to hide.

This is how we treat the idea scientifically rather than as consolation. We either earn support by surviving hard tests, or we publish failure.

3.5.11 Where we go next:

The endpoint is fixed: the ledger closes at *death of mind*. Next, 3.6 provides the technical spec—how we estimate momentary valence $V(t)$ (via a delta-based Hedonic Composite Index), integrate to *HCU*, and apply equivalence testing to adjudicate neutrality. It details procedures for abrupt closures and other confounds so failures, if any, are unmistakable.

3.6 Research Notes: The Ledger Integral and State–Change Formalism

These notes lay out the mathematical scaffolding for Chapter 3: how we represent the life ledger formally, how we estimate the latent affect process $F(t)$ (momentary net affect), how uncertainty propagates into the total $L(T)$, how we handle splits and merges quantitatively, and how we test neutrality with equivalence methods. The aim is to be concrete enough that an independent researcher could implement these analyses or challenge them. In short, our approach treats experience as a continuously evolving internal state: a state–change formalism rather than any static label. The Hedonic Composite Index (HCI) is our composite metric tracking this ongoing change, and integrating it over time (summing its moment-to-moment deltas) yields the cumulative life ledger.

3.6.1 Hidden state, observations, and channels

We model the momentary net affect $F(t)$ as a latent continuous-time state that is not directly observed but influences multiple measurable indicators. At any time t , $F(t) \in \mathbb{R}$ on an anchored HCI scale (0 = neutral; positive = pleasant; negative = unpleasant). This represents the hidden state of how good or bad the person feels at that moment.

We observe a set of channels: signals that carry information about $F(t)$. Let $Y_k(t)$ denote the observation from channel k at time t , for $k = 1, \dots, K$. Examples include self-reported mood ratings, heart-rate variability, electrodermal activity, facial-expression scores from video, EEG-derived features, and dream-content coding. These jointly form a multivariate time series of observations.

For channels that are roughly linear (or can be treated as linear in the range of interest), we use a Gaussian observation model:

$$Y_k(t) = \alpha_k + \beta_k F(t) + \varepsilon_k(t), \quad \varepsilon_k(t) \sim N(0, \sigma_k^2).$$

Here α_k is a baseline offset for channel k , β_k is a sensitivity (slope) for how changes in F are reflected in channel k , and ε_k is noise.

For binary or saturating channels, we use a generalized-link observation model:

$$g_k(E[Y_k(t) | F(t)]) = \alpha_k + \beta_k F(t),$$

where g_k is an appropriate link function (e.g., logit or probit for binary events; or a log/logistic transform for bounded measures that saturate). In these cases, stochasticity is captured by the assumed distribution rather than an additive noise term.

We incorporate within-person calibration: an individual i may have channel-specific parameters $\alpha_k(i)$ and $\beta_k(i)$. We estimate these via hierarchical (multilevel) models,

pooling information while allowing person-level differences. For identifiability, we constrain $\beta_k > 0$ for channels with known monotonic direction (e.g., more smiling implies higher F , so β for a smile detector is positive by definition). For ambiguous channels, we let the data determine the sign.

Conceptually, this is a confirmatory factor model with a single latent factor $F(t)$ at each time point. The Hedonic Composite Index (HCI) is essentially the resulting factor score, our estimate of $F(t)$ given all current observations. The next subsection describes how $F(t)$ evolves over time (the state-space dynamics that propagate this estimate forward).

One can read our definition $dL/dt = F(t)$ through the lens of active inference: in variational terms, valence corresponds to the sign of change in expected surprise. When prediction error decreases, experience trends positive; when prediction error increases, experience trends negative. This interpretation does not change our math. It provides one plausible computational story for how ordinary control loops could implement a constraint-like closure condition without invoking teleology. (This is a computational gloss consistent with our formalism, not a new assumption.)

We also use anchor tasks to calibrate units across people and channels. For example, we might employ a standardized cold-pressor pain task and a mild analgesic-relief task to define what “ ± 1.0 HCU” means. Specifically, we could define that a controlled increase in pain corresponds to -1 HCU and a matched controlled relief corresponds to $+1$ HCU. By doing this, we align different people’s scales. For instance, suppose the median change in a person’s HCI from a standard dose of IV morphine is $+X$; we could choose that magnitude to represent $+1.0$ HCU for that individual. The goal is for 1 HCU to have a concrete meaning tied to real experiences for everyone.

For example, we might define:

- $+1$ HCU as the net effect of a one-minute elevation in F equal to a “small positive” anchor experience.
- -1 HCU as the corresponding effect of a one-minute exposure to a symmetric “small negative” anchor.

A “small positive” anchor could be the mild mood boost from listening to a favorite piece of music or receiving a warm hug: noticeable, but not extreme. We would gather data to quantify that effect. Suppose the median self-reported mood increase from that experience corresponds to $+0.5$ on some internal scale; we would then label sustaining that magnitude for one minute as $+1$ HCU. Likewise, a “small negative” anchor might be a mild social rejection or losing a small amount of money: briefly upsetting but not

overwhelming. One minute at the matched negative magnitude would be defined as -1 HCU.

We likely use multiple anchor scenarios (e.g., one physical stimulus and one social stimulus) to build an anchor ladder, then fit a graded model so that the HCl scale is aligned across contexts.

Differential definition and integration: With anchors in place, the delta-HCI operationalizes the instantaneous net affect rate. On grid t_i ,

$$HCl(t_i) = \sum_k w_k \cdot \Delta z_k(t_i) \text{ (standardized first differences).}$$

The ledger over any interval $[a, b]$ is numerically integrated as:

$$HCU[a, b] = \sum_{m: t_m \in [a, b]} HCl(t_m) \cdot \Delta t_m \text{ (e.g., trapezoidal rule).}$$

By calibration, 1 HCU equals “one minute at HCl = +1” (and similarly for negative anchors), making $L(T) = \int_0^T F(t) dt$ interpretable in lived units.

This anchored approach ensures our units are not arbitrary. It also means the neutrality bound $\pm K$ can be set meaningfully in these units. For example, $K = 5$ HCU might correspond to the equivalent of about five small good experiences more than bad, a net difference considered trivial or within the noise.

In summary: at any time t we fuse multiple noisy observations (through the above observation models) to estimate $F(t)$ with uncertainty. The channels are treated as noisy linear or linked functions of $F(t)$, with parameters identified and partially person-specific (via hierarchical modeling). We constrain signs where appropriate so that the composite (HCl) remains interpretable and auditable.

3.6.2 Dynamics for $F(t)$

We posit that $F(t)$ (net affect) evolves as a stochastic process with certain properties: it tends to revert toward a baseline level but can be jolted by discrete events. A plausible model is a mean-reverting Ornstein–Uhlenbeck (OU) process with jumps:

$$dF(t) = -\kappa(F(t) - \mu(t)) dt + \sigma dW(t) + \sum_j \eta_j \delta(t - t_j).$$

Breaking that down:

- $\kappa > 0$ is the reversion rate: how quickly affect tends to pull back toward a moving baseline $\mu(t)$. If someone’s affect is above baseline, κ gently drags it down; if below baseline, κ nudges it up. This is a mood-homeostasis parameter, a loose spring restoring equilibrium.

- $\mu(t)$ is a slowly drifting baseline (which could be person-specific and time-varying, reflecting circadian rhythm or long-term adaptation). We can model $\mu(t)$ as a smooth function (e.g., a Gaussian Process prior) or as piecewise-constant with a prior on changes, capturing different set-points or trends across life stages.
- $\sigma dW(t)$ is a Brownian noise term (Wiener process) representing random fluctuations with volatility σ . This covers minor ups and downs that aren't driven by specific events: the background noise of mood.
- $\sum_j \eta_j \delta(t - t_j)$ represents discrete impulses at times t_j with magnitudes η_j . These are events that cause immediate jumps in $F(t)$, such as a sudden injury (large negative jump) or a joyous reunion (large positive jump). Here $\delta(\cdot)$ is the Dirac delta, signifying an instantaneous change at that moment.

We treat κ and σ as parameters to infer (possibly hierarchically: some people's affect may be more volatile, higher σ , or more "sticky," higher κ , than others). We give $\mu(t)$ a prior that encourages smoothness or slow change (to capture gradual adaptation or circadian effects). Jump magnitudes $\{\eta_j\}$ can be inferred from the data, potentially informed by context when known (e.g., if at time t_j the person got married, we might place a prior expecting an upward jump), or treated as latent events when unknown (the inference attributes large unexplained shifts to unmodeled impulses).

This stochastic differential equation (SDE) is continuous in time; in practice, we discretize it for computation or use filtering methods (unscented Kalman filters, particle filters, etc.) to estimate $F(t)$ given irregularly spaced data.

Why this model? It captures two key intuitions from affective science: (1) affect is autocorrelated and tends not to wander off to infinity (there's a regulatory mechanism keeping moods within bounds – the κ term pulling toward baseline), yet (2) significant life events can perturb mood abruptly (the jump term allows sudden changes). This aligns with the notion of a homeostatic mood system subject to perturbations: people have a general mood set-point, but life events knock them around that baseline.

We fit these dynamics using Bayesian filtering methods. The continuous OU process provides a tractable way to handle irregular self-report times and asynchronous data streams from wearables, by predicting how someone's affect evolves in the gaps and updating estimates when new observations arrive. In practice, the OU + jumps model lets us predict the trajectory of $\hat{F}(t)$ between observation points and then refine those predictions as data come in.

Having specified the dynamics, we can feed in our data and obtain an estimate $\hat{F}(t)$ (the posterior mean latent affect at time t) along with uncertainty (the posterior variance at

each t). That time-varying estimate *is* essentially the HCl as a function of time (with an uncertainty band). In other words, after combining all observations and accounting for dynamics, we get a continuous inferred trajectory of the person’s net affect.

3.6.3 From HCl to HCU: units and anchors

We define 1.0 Hedonic Composite Unit (HCU) to have a clear real-world interpretation using the anchor points described above. In practice, we scale $F(t)$ (initially expressed in arbitrary “HCl units” per unit time) such that:

- +1 HCU equals the effect of sustaining a “small positive” anchor experience for one minute.
- -1 HCU equals the effect of a symmetric “small negative” anchor for one minute.

For example, if a “small positive” anchor is the mild mood boost from a friendly conversation or a favorite piece of music, we determine—via model estimation or calibration tasks—what change in F corresponds to that experience. Suppose the data show that this produces approximately a +0.5 shift on a standardized internal scale; we label sustaining that magnitude for one minute as +1 HCU. Similarly, a “small negative” experience (e.g., a mild social slight or minor loss) might be defined as -1 HCU per minute at its matched intensity.

In practice, we use multiple anchors. For instance, we may define ± 1 HCU using a physical pain/relief pair (such as cold-water immersion versus nitrous oxide analgesia) and cross-check with a psychological pair (e.g., social inclusion versus exclusion). We then ensure that these anchors yield consistent scaling across contexts. This anchored calibration ensures that HCU units are interpretable, comparable across individuals, and grounded in lived experience rather than abstract scale values.

It also allows us to choose a meaningful neutrality band $\pm K$ in HCU units. For example, $K = 5$ HCU might represent a net difference equivalent to only a few minor positive experiences over negative ones—an imbalance small enough to count as practically negligible within our predefined tolerance.

(Section 3.6.1 introduced the anchors conceptually; here we operationalize them for quantitative analysis.)

3.6.4 The ledger integral and its estimator

For a single unified stream over its lifetime $[0, T]$, the true lifetime ledger is defined as the time integral of net affect:

$$L(T) = \int_0^T F(t) dt,$$

measured in HCU (since $F(t)$ is calibrated into HCU per unit time after anchoring).

State–Change formalism (mechanism to signal identity): Let $S(t) \in \mathbb{R}^d$ denote the latent drive-load vector (homeostatic, threat, social/meaning, comfort, and related components). Let $W \in \mathbb{R}^{1 \times d}$ be a fixed weighting row vector. Define the felt valence rate (our net affect integrand) by:

$$F(t) = V(t) = -W \cdot dS(t)/dt.$$

Pleasure corresponds to a net reduction in drives ($F(t) > 0$); pain corresponds to a net increase ($F(t) < 0$).

Closure implication: Because

$$L(T) = \int_0^T F(t) dt = -\int_0^T W \cdot (dS(t)/dt) dt = -W \cdot (S(T) - S(0)),$$

if the boundary condition holds that $S(0) = 0$ and $S(T) = 0$ (birth/death of mind at drive baseline), then $L(T) = 0$. The empirical task is therefore to estimate $F(t)$ throughout life and test whether the lifetime ledger $L(T)$ is statistically equivalent to zero within a preregistered ROPE.

Observation model (long form): At observation times t_i , channels $y_k(t_i)$ (affect EMA, physiology, sleep/dream affect, behavior, neural markers) are modeled as:

$$y_k(t_i) = h_k(S(t_i), \Delta S(t_i)/\Delta t, F(t_i)) + \varepsilon_k(t_i),$$

with $\varepsilon_k(t_i)$ zero-mean noise. This long-form model explicitly ties measured signals to state and state change, avoiding category errors.

Delta-based HCI (operational $V(t)$): On a fixed grid t_i , compute standardized first differences for each validated channel $\Delta z_k(t_i)$ and define:

$$HCI(t_i) = \sum_k w_k \cdot \Delta z_k(t_i).$$

Operationally, we treat $HCI(t_i)$ as our estimate of the momentary felt-balance rate $F(t_i)$ (units: HCU per unit time after calibration in §3.6.1 and later calibration details in Chapter 7)

Neutrality test (equivalence form): Choose a neutrality bound $K > 0$ (ROPE). Test at the end of mind:

$$H_0: L(T) \leq -K \text{ or } L(T) \geq K \text{ (non-neutral)}$$

$$H_1: -K < L(T) < K \text{ (neutral within ROPE)}.$$

Report both $L(T)$ and its uncertainty (see covariance integral below) and conduct TOST-style equivalence.

We cannot observe $F(t)$ exactly, but after fitting our state-space model we have a filtered posterior distribution for the latent path given all observations up to T : $p(F(\cdot) | Y_0:T)$. We compute $\hat{L}(T)$, an estimate of $L(T)$, along with an uncertainty interval.

One approach uses the posterior mean path $\hat{F}(t)$ and integrates it numerically. For example, partition $[0, T]$ into an adaptive grid t_m and compute a Riemann sum:

$$\hat{L}(T) = \sum_{m=1}^M \hat{F}(t_m) \Delta t_m,$$

refining the grid where F changes rapidly. This yields a point estimate of the ledger.

We also need uncertainty in $L(T)$. We can approximate it by integrating posterior covariance. By the delta method, roughly:

$$\text{Var}[L(T)] = \sum_m \sum_n \text{Cov}(\hat{F}(t_m), \hat{F}(t_n)) \Delta t_m \Delta t_n.$$

In practice, we approximate this double sum using the posterior covariance matrix of the filtered (or smoothed) \hat{F} values at our grid points (e.g., from a Kalman smoother).

If directly computing covariance is intractable (especially with nonlinear or particle filtering methods), a sampling approach is often simpler. After fitting the model, draw S samples from the joint posterior of the entire latent path $F(t)$ (e.g., via a particle smoother, or forward–backward sampling in a Kalman smoother for linear-Gaussian models). For each sample s , compute:

$$L^s(T) = \int_0^T F^s(t) dt,$$

via a discrete sum (or by integrating under a piecewise-linear assumption between sampled points). The values $\{L^s(T)\}$ represent draws from the posterior distribution of the ledger. From them, we calculate the posterior mean, variance, and a credible interval for $L(T)$. This Monte Carlo approach is conceptually straightforward and often easier to implement, since it leverages standard latent-path sampling algorithms.

Throughout these calculations, we account for missing data by letting uncertainty expand during gaps; the state-space filter already does this (fewer observations \rightarrow wider posterior for $F(t)$). Additionally, when needed we incorporate covariates for missingness (e.g., a “not measured” indicator influencing the state; see Section 3.6.13) to handle systematic dropout.

In summary: we integrate the latent net-affect trajectory to get a cumulative ledger, and we quantify our confidence in that integral.

3.6.5 Branching and fusion of ledgers

For cases where the stream branches (see identity split rules in Section 3.4.3), we handle the integration accordingly. Suppose we have B concurrent branches during some interval. The formal expression for the total ledger at time T is:

$$L(T) = \sum_{\{b=1\}}^B \int_{\{I_b\}} F_b(t) dt,$$

where I_b is the time interval during which branch b is active. If branches run strictly in parallel (two conscious centers at the same time), those intervals I_b overlap; if they alternate (no true parallel consciousness, just switching), then the intervals tile the timeline without overlap.

In words: integrate each branch's affect over the period it was active, then sum. If two branches run strictly concurrently (e.g., a split-brain patient with two independent streams for a period), we sum their integrals because experiences were happening simultaneously and separately in each branch.

We assume branch processes $F_b(t)$ are independent during true concurrency (by definition, if the person is truly split into separate awareness streams, those streams do not directly interact). Thus, covariance between different branch contributions is taken as zero while they are concurrent. If branches alternate in time (so only one is “on” at once), the formula reduces to integrating one process at a time in sequence, which is equivalent to integrating the unified timeline under the split rules.

When branches fuse back together (Section 3.4.4 covers identity fusion), we carry forward the sum of their integrals as the continuing ledger. For uncertainty: if branches are independent, variances add over disjoint intervals. If there is interaction or later correlation between branches (e.g., upon fusion one stream's memories become accessible to the other, correlating estimation errors), we account for covariance where estimable. Our default assumption is negligible covariance at fusion unless the measurement model implies otherwise.

Importantly, we apply no moral weighting or arbitrary discounting. It is arithmetic addition of HCU. We also ensure no double-counting within a branch: each branch's integral covers its own subjective experience domain. If two branches were truly concurrent, then yes, the same clock-minute can contribute twice (once per mind), because two independent experience-minutes occurred. This is rare and largely theoretical, but the rule is explicit: if there are truly two independent centers of awareness in one body, we treat them as separate individuals for that duration.

In most normal cases, such fragmentation either does not occur or is brief. But we have a method: track each ledger separately during any split, then sum them if the streams rejoin.

3.6.6 Equivalence testing at closure

At death-of-mind time T (closure of the ledger), we have our estimated ledger $\hat{L}(T)$ and a posterior distribution for it as described above. To decide whether the lifetime ledger is effectively zero (balanced), we perform an equivalence test (either frequentist Two One-Sided Tests or a Bayesian interval test) against the null hypothesis that there is a meaningful imbalance beyond $\pm K$ HCU.

In a frequentist approach, we set up two one-sided tests:

- $H_1: \hat{L}(T) > -K$ (i.e. the ledger is not too far negative)
- $H_2: \hat{L}(T) < +K$ (i.e. the ledger is not too far positive)

This is a non-inferiority style test on both sides. We use the 90% confidence interval of $\hat{L}(T)$ (for a 5% equivalence criterion) and check whether that interval lies entirely within $(-K, +K)$. If yes, we declare Neutral (equivalence achieved). If not, and the interval lies entirely outside the bounds in one direction, we declare Non-neutral (a clear imbalance). If it overlaps the boundary, we declare Inconclusive.

In a Bayesian version, we compute the posterior probability that $L(T)$ is within $\pm K$: $\Pr(-K < L(T) < K | \text{data})$. If that probability ≥ 0.95 , we label Neutral; if most posterior mass lies outside the $\pm K$ band, label Non-neutral; otherwise, label Inconclusive.

We report both types of analysis. We also preregister a sensitivity analysis: repeat equivalence with a stricter bound (e.g., $K/2$) and a looser bound (e.g., $2K$) and report whether conclusions are robust. The primary decision rule remains the preregistered K .

Finally, we categorize the result for each stream at closure as Neutral, Non-neutral, or Inconclusive, following the criteria set in Section 3.4.17.

3.6.7 Horizon scaling predictors

To test the prediction that as one's horizon shrinks (as the end of life approaches), compensatory impulses grow stronger in the opposite direction (see Section 3.5.5), we introduce an explicit horizon variable H_t , estimating expected remaining time at time t (from clinical prognosis or actuarial models, updated as appropriate). We then examine magnitudes of event impulses η_j in relation to both recent ledger state and horizon.

For instance, we can fit a regression for positive-event impulses in our SDE model such as:

$$\eta_j \sim \mathcal{N}(\gamma_0 + \gamma_1 \Delta^- L(t_j^-) \cdot h(H(t_j)), \tau^2).$$

Here $\Delta^- L(t_j^-)$ is the recent negative ledger imbalance just before that event, and $h(H)$ is a decreasing function of remaining time (e.g., $1/H$). The key prediction is $\gamma_1 > 0$.

This means that if there is a large recent negative deficit and the horizon is short, compensatory positive impulses tend to be larger on average: an interaction effect in which a short horizon amplifies compensation for prior suffering.

We test $\gamma_1 > 0$ at the population level, for example using hierarchical models across individuals. We include appropriate covariates (baseline drift, long-term trends, diagnosis, medication, and other confounds) to isolate this effect.

3.6.8 QS-residuals in control circuits

We seek evidence of the Queue System (QS) via neurobehavioral data by testing for a QS-residual signal in cognitive-control circuitry. For example, in an experiment or naturalistic data, we might measure a control-related neural or behavioral signal $A(t)$ (activity in rIFG or ACC, or a composite measure of inhibition/control effort).

We fit a regression model incorporating known influences and our QS term:

$$A(t) = \beta_0 + \beta_1 U(t) + \beta_2 C(t) + \beta_3 \text{Arousal}(t) + \beta_4 \Phi(t) + \varepsilon(t).$$

Here $U(t)$ is utility or reward value, $C(t)$ is conflict, $\text{Arousal}(t)$ is general arousal (e.g., pupil diameter or heart rate), and $\Phi(t)$ is our computed feasibility-of-compensation metric, a function of ledger imbalance and remaining horizon.

In this regression, β_4 captures the effect of compensation pressure Φ on control signal A . The prediction is $\beta_4 > 0$: higher compensability systematically recruits control circuitry and biases policy selection toward repair options after accounting for utility, conflict, and arousal. If β_4 consistently equals 0 (no effect of Φ) under adequate designs, we have not detected QS activity in these signals.

We account for multiplicity and apply preregistered correction procedures where relevant. (Chapter 5 expands on related experiments.)

In building this model, we account for multiple comparisons and apply appropriate corrections (since we might test this across various tasks or regions of interest). If we find that β_4 is significantly > 0 across many tasks or datasets, it supports the presence of a QS mechanism influencing control processes. If β_4 consistently equals 0 (no effect of Φ), then we haven't detected QS activity in these signals.

3.6.9 Dream counterweights

To test dream compensation (introduced conceptually in Section 3.5.5), we analyze sequences of days and subsequent nights' dreams. For each sleep period, we quantify dream affect D_d for night d , for example by scoring dream reports and, where appropriate, using physiological proxies (e.g., sleep-stage measures and EEG features that correlate with affective tone).

We might use a model like:

$$D_d = \alpha + \rho D_{d-1} + \lambda \Delta^- L_{(d-1)} + \xi_d,$$

where $\Delta^- L_{\{d-1\}}$ is the prior day's net affect (negative when the day's net affect was negative). We include an AR(1) term ρ because dream affect is likely autocorrelated night-to-night (baseline tendencies in dream emotion), and we include relevant controls (e.g., sleep architecture measures such as % REM sleep) since these can influence dream emotionality independently.

The key parameter is λ , and we expect $\lambda > 0$. This would mean that if yesterday was more negative, then tonight's dream affect is more positive, all else equal: an emotional inversion. A positive λ would indicate systematic counterweights (compensatory positive dream affect after negative days). We formally test $\lambda > 0$. If λ is around 0 (or negative), we find no evidence of the compensatory pattern (or evidence of congruence where dreams echo negativity, which would contradict the counterweight signature).

3.6.10 Variance compression near closure

We next examine whether variance compression occurs as the ledger closes. Using longitudinal data from cohorts approaching end of life (or a proxy such as very advanced age, though known horizon is preferable), we examine the between-person variance in partial ledger totals as a function of proximity to death.

One formulation treats $\text{Var}[L(t)]$ as a function $f(t)$ that should decrease as $t \rightarrow T$. Alternatively, we compare distributions of ledger totals at fixed offsets before death (e.g., 12 months, 6 months, 3 months, and near T). The prediction is $\text{Var}[L(t)]$ across individuals decreases as $t \rightarrow T$, after accounting for measurement noise.

We must adjust for data quality near death: fewer measurements and higher uncertainty can distort apparent variance. We therefore account for known uncertainty in each individual's ledger estimate (e.g., via confidence intervals) to distinguish true outcome variability from measurement uncertainty.

3.6.11 Cross-person linking and invariance

To compare or aggregate ledgers across people, especially across cultures or languages (recall invariance requirements in Section 3.4.12), we require measurement invariance of the model in Section 3.6.1. In practice, we fit multi-group latent-factor models to test invariance.

We proceed in standard steps:

- Configural invariance: does the same loading structure hold in each group?
- Metric invariance: constrain loadings β_k to be equal across groups and assess whether fit deteriorates substantially (e.g., using a criterion such as $\Delta\text{CFI} < 0.01$).
- Scalar invariance: further constrain intercepts/thresholds; if full scalar invariance fails, allow partial invariance or use alignment methods, then propagate the linking uncertainty.

If we cannot achieve at least metric invariance, we do not pool across groups. We analyze within groups only. Assuming invariance holds, we may also use physical anchors (identical pain stimuli across cultures where ethically appropriate) as additional linking points, under the assumption that certain responses are biologically comparable enough to support scale alignment.

If invariance checks out, we proceed to pool results across people, adding extra uncertainty when partial invariance or alignment is used. This is how we ensure that “1 HCU” means approximately the same amount of felt change for person A as for person B, within stated error bounds. Without invariance, cross-person ledger differences could be artifacts of incomparable scales.

3.6.12 Priors, computation, and diagnostics

We adopt reasonable priors for model parameters to ensure identifiability and avoid overfitting:

- Dynamics priors can be weakly informative (e.g., $\kappa \sim \Gamma(2,1)$ under an appropriate timescale parameterization; σ with a half-Cauchy or half-normal prior). Hierarchical priors on channel gains β_k allow individual differences but regularize extremes. If we infer impulses $\{\eta_j\}$ from data, sparsity-inducing priors (spike-and-slab, horseshoe) can prevent overfitting noise as major events.
- Hierarchical priors on channel gains β_k allow individual differences but regularize extremes. If we infer impulses $\{\eta_j\}$ from data, sparsity-inducing priors (spike-and-slab, horseshoe) can prevent overfitting noise as major events.

- Computation: We use Bayesian inference (e.g., HMC via Stan/NumPyro) after discretization or with suitable approximations, or particle methods for nonlinear/non-Gaussian components. We use model comparison metrics such as LOO or WAIC to compare variants (e.g., with versus without QS terms), and we monitor convergence diagnostics (e.g., $\hat{R} < 1.01$ and adequate effective sample size).
- Diagnostics: We perform posterior predictive checks (can the model reproduce the distribution and autocorrelation of affect changes) and simulation-based calibration (recover known parameters from synthetic datasets) to detect estimation bias or implementation errors.

Handling missing data and non-ignorable dropout If missingness is not random (e.g., participants skip reports when feeling particularly bad), we adjust for it to avoid bias (e.g., underestimating suffering). One approach is to include an “availability” sub-model in the state-space framework, treating missingness as another output that depends on $F(t)$. For example:

$$\Pr(\text{observation at time } t \text{ is missing}) = f(F(t)),$$

for an appropriate link function f . Alternatively, we perform sensitivity analyses under MAR versus MNAR assumptions, including worst-case missingness bounds, and report how neutrality conclusions shift.

If end-of-life data are missing because participants are unresponsive or too ill to report, we rely more heavily on physiological proxies where possible and explicitly flag those cases. What we do not do is ignore missing data or carry the last observation forward indiscriminately. We integrate missingness into the model and propagate uncertainty honestly. If uncertainty remains too large at T , we classify the case as Inconclusive.

3.6.13 Preanalysis plan (minimal)

We outline the minimum prespecified steps to keep testing transparent:

- Primary endpoint: the lifetime ledger $L(T)$ with a 95% confidence interval (frequentist) or credible interval (Bayesian), and an equivalence test against $\pm K$ HCU (K defined from anchors; e.g., K might correspond to about 0.15 SD of the composite scale, depending on calibration).
- Primary signatures to evaluate:
 - γ_1 – (horizon scaling coefficient, Section 3.6.7)
 - β_4 – (QS-residual coefficient, Section 3.6.8)
 - λ – (dream inversion coefficient, Section 3.6.9)

- Slope of $\text{Var}[L(t)]$ as t approaches T (variance compression, Section 3.6.10)

We will evaluate each of these against 0 (or against the null of no effect, equivalently). These correspond to the key predicted signatures of the Law of Fairness.

- Covariates/stratifiers: age, sex, diagnosis, medication usage, socioeconomic variables, culture/language, and device/hardware differences. These enter as covariates or via partial pooling to reduce confounding of the signatures.
- Multiplicity: Because we have multiple primary tests, we adjust for multiple comparisons. In frequentist analyses, we may use Holm–Bonferroni to control family-wise error. In Bayesian analyses, we report joint posterior quantities transparently; hierarchical modeling can reduce false positives without “fishing.”
- Stopping rules: If interim checks show futility or participant burden rises beyond ethical thresholds, we pause and reassess. For observational datasets, “stopping” mainly concerns analysis staging; rules are prespecified to prevent goalpost shifts.

3.6.14 Failure modes and what they imply

We prespecify how we interpret key failure patterns:

- If posterior variance of $L(T)$ remains large in most cases: the meter is inadequate; improve measurement before drawing strong conclusions.
- If measurement invariance cannot be achieved across key groups: claims must be limited to within-group comparisons; universality cannot be supported empirically.
- If γ_1 , β_4 , and λ are null with tight intervals around 0: mechanism signatures are absent; the compensatory architecture is not supported.
- If frequent terminal non-neutralities occur beyond $\pm K$: the Law fails for those streams. If failures cluster under “inhumane conditions,” a contingent “only under humane conditions” claim would not be a universal law of nature.

We commit that if these failure conditions appear and persist with adequate power and rigorous analysis, we will declare the Law unsupported or false in those respects. We will not redefine terms or retreat to ad hoc excuses.

3.6.15 Reproducibility kit (what labs should release)

We advocate transparency and reproducibility, especially because this claim is extraordinary. Any lab testing the Law of Fairness should share at least:

- Simulated datasets with known ground truth for $F(t)$ (including known quirks/noise characteristics) to validate pipelines.
- Full code for filters, inference procedures, and preregistered analyses (model code and scripts).
- Anchor task specifications and calibration tables (exact mappings from tasks to HCU).
- A “closure packet” template per participant: Unity Index time series near the end; cumulative ledger with intervals; key annotations, appropriately de-identified.
- A negative-results archive logging null findings and failed replications to reduce publication bias.

We emphasize the last point because if dozens of attempts fail but only one success is published, the record becomes falsely positive. Transparency about null findings is essential.

3.6.16 Where we go next:

With the ledger formalized and grounded in data, we zoom out in Chapter 4 to place the Law in its proper category. We insist it functions as a constraint, a guardrail, not as a goal or intention of any system. The lifetime ledger $L(T)$ is not a vague notion; it is a calculable quantity with uncertainty bounds. By combining a principled latent-state model, anchored units, predeclared boundaries, and proper equivalence tests, we make the Law of Fairness into a claim that one can actually attempt to falsify or support with data. This is crucial. Otherwise, it is just philosophy.

Chapter 4 — Constraint, Not Purpose

A hospice nurse quietly calls it a “miracle.” After years of estrangement, her patient Diego has welcomed his daughter back into his life, finding peace just days before he passes away. The reconciliation seemed to come out of nowhere, as if some invisible force wanted father and daughter to have closure. Stories like this feel intentional, as if the universe or fate were guiding events toward a fair ending. But is there another way to see it? Perhaps Diego’s own mind, sensing the horizon of life drawing near, shifted his priorities and nudged him toward making amends. What looks like a cosmic plan might instead be an internal adjustment: a late-stage course correction emerging from ordinary psychological processes. In this chapter, we explore the idea that if fairness emerges in lives, it does so not by grand design or “meant to be,” but by many small constraints quietly shaping what is possible.

The Law of Fairness, if true, would act as a guardrail, not a chauffeur. It would limit the range of possible outcomes, preventing a life from ending in imbalance, without actively steering each life toward a predestined “happy ending.” This distinction between constraint and purpose is crucial. The Law does not imply any conscious intent in the universe, nor that individuals deliberately aim for balance. Instead, irredeemably unfair life trajectories would be naturally pruned by the system’s dynamics. In plain terms, lives do not unfold under a cosmic plan to even things out; if the Law holds, terminal imbalances cannot occur because the structure of the system will not permit them.

Viewing fairness as a passive constraint makes it a scientific question rather than a spiritual one. Explaining the fairness pattern through known mechanisms such as adaptation, feedback loops, and social support networks is more convincing than invoking miracles or cosmic justice. In science, many small forces can add up to a large regularity. A law built from ordinary processes is sturdier than a belief in magical intervention. This constraint-based view avoids mystical thinking and generates concrete predictions. If a law is truly in effect, it should leave signatures in behavior and biology, some of which were outlined in Chapter 3. By focusing on constraints, we can predict surges of reconciliation or contentment as time grows short without ever saying the universe “wants” it.

We now situate the Law of Fairness within broader accounts of natural law. Philosophers of science distinguish between Governing laws, which actively compel outcomes, and Best-System laws, which summarize regularities without implying force or intent. Which is LoF? On one view, the fairness constraint would be a structural rule embedded in the dynamics of life trajectories, analogous to a conservation law. On another, it would be the most economical summary of converging regulatory processes. In practice, we lean

toward the latter interpretation with a constraint emphasis: LoF is treated as a working hypothesis that an underlying regulatory principle operates without purpose or agency. We search for its signatures, but we do not posit a cosmic actor.

Because we adopt a no-purpose stance, our language remains disciplined. We avoid phrases such as “nature wants to balance things.” Instead, we speak of probabilities, constraints, and weighted options. Rather than “Diego’s destiny required reconciliation,” we would say that, given his ledger state and shortening horizon, reconciliation became increasingly feasible and salient. Translating teleological language into mechanistic language is deliberate. It keeps the theory anchored in neuropsychology and decision dynamics rather than metaphor. Removing purpose does not weaken the claim; it strengthens its testability. A constraint without intention is easier to measure because it reduces to patterned cause and effect.

What you’ll get from this Chapter:

- Constraint vs. purpose (made clear): You will understand the difference between a law that bounds possible outcomes and one that implies goals. If the Law of Fairness is true, lives unfold within strict guardrails that forbid extreme terminal imbalances, not under guidance from an intentional force.
- Why constraints beat miracles: We examine why constraint-based explanations grounded in feedback and opportunity structure are scientifically stronger than narratives of intention or fate. A constraint-based account respects causal closure, generalizes across cases, yields quantitative predictions, and clearly specifies what would count as failure.
- No teleology allowed: We show how apparently goal-directed patterns, such as people finding peace near the end of life, can be described in mechanistic terms. Statements like “it was meant to be” can be reframed as “given the circumstances, that outcome became more feasible.” This discipline keeps the claim biologically grounded.
- Lawhood in perspective: We place LoF within debates about Best-System and Governing laws. We explain why, for practical purposes, we treat LoF as a Best-System style constraint hypothesis and how that framing strengthens experimental design and interpretation.
- Research notes – stopping time and signals: In the technical sections, we formalize “death of mind” as a stopping time for the ledger process and sketch how a shrinking horizon H can alter weighting dynamics, raising an internal shadow price λ_t as closure approaches. We clarify the limits of the Optional

Stopping Theorem in this setting and describe safeguards such as pre-registered endpoints and dropout controls to prevent statistical artifacts.

- Fail patterns highlighted: Throughout this chapter and the book, we identify specific fail patterns that would break the Law. By the end of Chapter 4, the central fail condition is explicit: well-tracked lives that end with large, uncorrected hedonic imbalances despite open channels for compensation. A theory that names its own possible refutation is operating scientifically, not offering comfort.

Subsections in this Chapter:

- **4.1 Guardrails vs. Goals** – Uses a concrete picture to distinguish constraints from aims. We define admissible sets and show how a guardrail removes extreme, non-balancing paths without “trying” to help anyone.
- **4.2 Why Constraints Beat Miracles** – Lays out the scientific case for distributed lawful processes over grand purposes. We show how ordinary mechanisms such as adaptation and feedback produce the signatures we test for, and why teleology adds nothing predictive
- **4.3 Lawhood: Best System vs. Governing Law** – Places the claim within philosophy of science. We explain what it would mean for LoF to be a summary of regularities rather than a force that pushes events, and why that framing sharpens the tests.
- **4.4 Language Discipline: Staying Non-Teleological** – Codifies permissible language (feasibility, selection, horizons) and language we avoid (“wants,” “deserves”). Clear wording guards against muddled inference and keeps the claim audit-ready.
- **4.5 Research Notes: Optional Stopping and Regularity** – Addresses statistical traps such as peeking, post hoc slicing, and irregular sampling, and specifies the diagnostics required to keep analyses decision-grade.

Where we go next:

We begin with 4.1. Keep the image of a mountain road in mind: constraints prevent catastrophic paths without aiming at any destination. That image will carry through the rest of the chapter.

4.1 Guardrails vs. Steering

A guardrail changes which paths are possible; steering selects among those paths. If the Law of Fairness holds, it behaves like a guardrail. It does not drive choices moment to moment. Instead, it bounds the space of possible life trajectories so that any ending incompatible with neutral closure cannot occur.

4.1.1 The mountain road picture

You still drive your own car; weather, traffic, and skill all still matter. The guardrail never grabs the wheel; it only prevents certain crashes. Likewise, the proposed Queue System (QS) does not insert new desires. It subtly weights familiar processes such as attention, valuation, and inhibition so that some options never gain traction or cannot be sustained. QS is a guardrail on the mountain road: you remain the driver, but some disastrous turns are removed.

4.1.2 What a guardrail actually does (and does not do)

Does:

- Prune options whose downstream consequences would make neutral closure implausible, given the current ledger state and remaining horizon.
- Amplify the salience of options that preserve or restore compensability, especially as the horizon shrinks.
- Modulate control thresholds (attention, inhibition, persistence) so that certain impulses feel “not this” or “now this” at the right moments.

Does *not*:

- That any external agent is steering events toward a goal. The constraint operates through ordinary psychological and neural processes.
- That daily life will be comfortable or fair. The Law concerns closure at T, not the smoothness of the path. Hard but compensable paths remain admissible.
- That the menu must remain broad. In extreme circumstances it may narrow sharply; what matters is whether neutral closure remains feasible.
- That preferences or beliefs cannot change. The model allows shifts in valuation and priority; it rejects only teleological interpretations of those shifts as cosmic purpose.

Agency remains: Even with a smaller menu of options, it is still your menu. Among the admissible options, you still steer. Ethics and law routinely acknowledge real constraints (duress, incapacity) without erasing personal agency. The QS constraint is not duress; it

is a background weighting of feasibility and salience. Responsibility for choices remains with the individual, along with the human work of judgment and care.

4.1.3 The admissible set, sketched formally

Let $\mathcal{U}(t)$ be the set of all thinkable actions at time t . Given the current ledger $L(t)$ and remaining horizon H_t , the admissible set $\mathcal{A}(t)$ consists of those actions $u \in \mathcal{U}(t)$ for which

$$\Pr[L(T) \in [-K, K] | L(t), H_t, u] \geq 1 - \varepsilon.$$

Formally:

$$\mathcal{A}(t) = \{ u \in \mathcal{U}(t) : \Pr[L(T) \in [-K, K] | L(t), H_t, u] \geq 1 - \varepsilon \}.$$

In words, $\mathcal{A}(t)$ contains actions that preserve a high probability of neutral closure. The QS weight $\omega(u; t)$ increases for actions that maintain this probability and decreases for those that reduce it. As the horizon shortens, the shadow price on non-compensability rises; $\mathcal{A}(t)$ contracts and weighting sharpens.

From the first-person perspective, this does not feel like command. It feels like certain ideas never arise or lose motivational force, while more reparative options feel timely and compelling.

4.1.4 Inside view: somatic tilts, not puppet strings

You feel a tug to call your sister, hesitate before sending an angry email, accept help more easily, or suddenly crave music or a walk. These are shifts in valuation and inhibition within decision circuits such as insula, vmPFC, ACC, and rIFG. They are leanings, not commands. If an action lies outside $\mathcal{A}(t)$, sustaining it becomes cognitively unstable. The chain of thoughts collapses before execution.

4.1.5 Outside view: a residual we can test

From an external perspective, neural activation $A(t)$ in control and valuation hubs can be modeled as:

$$A(t) = f(\text{utility}, \text{conflict}, \text{arousal}) + \beta \Phi(\hat{L}(t), H_t) + \varepsilon.$$

Here $\Phi(\hat{L}, H)$ captures feasibility of compensation given current ledger state and horizon. The prediction is $\beta > 0$ in regions such as rIFG, ACC, and vmPFC after controlling for utility, conflict, and arousal. This residual reflects systematic weighting within known circuitry, not a new force.

4.1.6 Everyday examples

These ordinary scenarios illustrate how the guardrail might manifest in daily life:

- The text unsent: The immediate utility of venting your anger is high, conflict level moderate, arousal high. But in a short-horizon context (say, it is late at night and the relationship is fragile), $\Phi < 0$ (sending the angry text would reduce the probability of later repair). Without you realizing it, your rIFG/ACC (inhibitory control regions) quietly hit the brakes; the impulse to send the text fades. You wake up glad you did not send it.
- The nap you suddenly “need”: After weeks of sleep deficit, $\Phi > 0$ for taking a rest; catching up on sleep would help rebalance your affective ledger. You feel a strong bodily “yes” to lying down despite your to-do list. (The guardrail does not care that you have emails to write; a compensatory nap is admissible and thus feels especially right.)
- The late reconciliation: As the horizon shortens (due to illness or advanced age), options that could close long-open ledgers through forgiveness or connection become more salient. Someone finds that apologizing or reaching out to an estranged loved one “suddenly feels easier” or urgently important; the words come more readily than they might have years earlier.

Each of these examples is completely mundane, which is exactly the point. The guardrails operate through ordinary psychological routes (shifts in attention, inhibition, “gut feelings”), not through dramatic voices from the sky.

In short, the QS “guardrails” often align with common-sense good habits. Everyday practices like getting enough sleep, exercising, or pausing before reacting in anger serve as informal guardrails; they keep us away from extreme emotional swings much as the QS would. Likewise, communities and cultures have developed their own guardrails (traditions of rest, forgiveness rituals, moderation in indulgence) that echo this natural constraint. (We will revisit these in Chapter 21, where daily routines, social support, and spiritual practices are treated as a “ledger gym” for maintaining balance.) The key is that none of these habits override your free will or magically “steer” you; they simply make it easier to stay within safe emotional bounds, consistent with the Law’s role as a background constraint.

4.1.7 How agency and responsibility survive

A constraint may shrink your menu of choices, but a smaller menu is still a menu. You remain answerable for which option you pick within $\mathcal{A}(t)$. (In human society we already recognize that some constraints on choice exist—for example, under extreme duress or

mental incapacity—without assuming that the person has zero agency. In our case, the QS constraint is gentle background weighting, far from the level of coercion.) So the presence of a fairness guardrail does not absolve anyone of responsibility. The Law of Fairness is not a fatalistic excuse to let harm happen (“destiny will balance it”). Normal ethical duties of care and judgment remain intact for every admissible choice. In short, the QS may shift the odds, but it does not choose for us; responsibility remains with the chooser.

4.1.8 What would falsify the guardrail picture

The guardrail hypothesis (QS) makes several predictions that could be proven false. Any of the following, if empirically observed and replicated, would seriously undermine or refute the idea:

- No “QS-residual” at all: Careful studies find zero evidence of any residual control-signal bias after accounting for utility, conflict, arousal, etc. (Nothing in rIFG/ACC/vmPFC correlates with $\Phi(\hat{L}, H)$ as predicted.)
- No horizon interaction: Compensatory options do not gain additional weight as the time horizon shrinks; even in end-of-life or other “last chance” contexts, no systematic uptick in balancing behavior occurs.
- Admissible-set leakage: We observe, repeatedly, life trajectories that are clearly non-compensable (trajectories that accumulate large negative ledgers with no feasible recovery) yet they proceed unimpeded to an imbalanced end, despite intact emotional and cognitive channels for compensation.
- An equally good rival model: A competing model with no constraint (for example, standard adaptation plus randomness) reproduces terminal neutrality and all purported QS dynamics just as well as our model with the constraint does.

Any one of the above findings, if robust, would force reconsideration of the Law. We would either fall back to explaining “balance” as a mere tendency of various processes (without a strict closure constraint) or abandon the Law of Fairness altogether. A theory that can fail these tests is participating in science; if it survives them, it gains credibility.

Fail Pattern (Terminal Imbalance):

Imagine researchers identify a substantial number of individuals who, despite having intact channels for emotional compensation (supportive social networks, no neurological deficits, available relief), end their lives with strongly negative or strongly positive total ledgers far outside the neutral bounds. If such terminal imbalances are well measured and persist without counterbalancing events, this would directly violate the

Law of Fairness. One clear case of an irredeemably unbalanced lifetime—replicated across contexts—would invalidate the guardrail constraint.

4.1.9 Why this framing helps the rest of the book

Thinking in terms of guardrails turns fuzzy metaphysical ideas into engineering questions. Instead of asking “does the universe care about fairness?”, we ask: Where are the weights applied in the mind? How large are they? How do they scale as the horizon shortens? Which empirical signatures distinguish a guardrail mechanism from ordinary adaptation? These are concrete questions that lead to measurements, and measurements lead to claims we can attempt to falsify. In the next section, we carry this mindset forward.

4.1.10 Where we go next:

With constraints distinguished from goals, 4.2 shows why this framing earns scientific trust. We trade miracle stories for regularities and derive predictions that later chapters can attempt to break.

4.2 Why Constraints Beat Miracles

A scientific explanation earns trust only if it respects causal closure, generalizes across cases, predicts distinctive signatures, and fails cleanly when it is wrong. A constraint-based account can do all of these. A miracle-based or teleological story cannot. If the Law of Fairness is real, it should present itself as a constraint on admissible life histories, not as a hidden hand that occasionally reaches in to override physics or “cheat” reality.

4.2.1 Causal closure stays intact

A miracle narrative introduces forces or intentions that break the known causal dynamics of life. By contrast, a constraint narrative leaves all ordinary dynamics in place; it simply bounds the set of trajectories those dynamics can realize. The organism’s existing machinery (interoception, valuation, control, learning, etc.) does all the work; the Law of Fairness, on this view, is a rule about which long-run outcomes are admissible. No extra push or mysterious energy is invoked. Nothing leaves or enters the ledger except what the organism can already feel and do through normal psychological processes.

Why this matters: A closed, naturalistic system can be simulated, perturbed, and replicated experimentally. If an explanation requires literal exceptions to normal causality, then it exits the domain of science. By insisting on causal closure, we ensure the claim can be tested with standard scientific tools.

4.2.2 Scalability across lives and contexts

Miracle stories tend to “solve” one case at a time; each balance becomes a one-off event. Constraint stories generalize. Once we specify what is bounded (terminal imbalance) and where the weights are applied (control and valuation hubs), we can test for the same signatures across settings: hospice wards, dorm rooms, disaster zones, or deep forests. The code path is the same; only parameters differ (horizon length, available compensation channels, health status). This creates a research program rather than a collection of anecdotes. A miracle story, by contrast, rarely scales beyond the anecdote.

4.2.3 Parsimony without emptiness

Constraint-based explanations are parsimonious. With one core rule (neutral closure by end-of-life within $\pm K$) and one implementation style (weight shifts in existing neural circuits), we replace a grab bag of ad hoc “just-so” fixes that would otherwise be needed to explain balance in each scenario. The account is not so general as to explain nothing; it makes risky predictions. The Law survives only if we observe the distinctive signatures it predicts (dream counterweights, horizon effects, QS-residuals, variance compression). These patterns are not trivial consequences of generic adaptation or

coincidence. In science, a simple hypothesis that makes precise, falsifiable predictions is stronger than a complex hypothesis that can accommodate anything. LoF is simple but not empty; it stands only if the predicted effects appear.

4.2.4 Falsifiability is built in

Miracle stories lack clear failure conditions; if the expected “miracle” does not occur, one can always claim delay or metaphor. A constraint story requires failure conditions. We pre-declare equivalence bounds for neutrality ($\pm K$ in HCU), set unity thresholds for ledger accrual, and specify null hypotheses for each predicted signature. If a well-tracked life ends with a ledger outside $\pm K$ and measurement is valid, the account fails for that stream. If the QS-residual vanishes after controlling for known factors, the mechanism fails. This asymmetry—making positive predictions while committing to crisp null outcomes—is what makes the claim scientific rather than rhetorical.

4.2.5 Robustness to noise and diversity

Real lives are heterogeneous. A constraint mechanism is robust because it operates through weights and feedback loops, not precise scripts. Different people, cultures, and species can satisfy the same guardrail constraint via different routes: one person finds reconciliation, another finds art, another finds solace in sleep or nature. The end result (neutral closure) is achieved through heterogeneity of micro-experiences. Miracles are brittle; if the special event does not occur, the balance collapses. A weight-based guardrail tolerates noise, missing data, and individual differences while still producing aggregate signatures. The constraint does not require identical lives; it requires convergence at closure.

4.2.6 Compatibility with agency

A miracle framing can undermine agency; if destiny ensures balance, individual choice becomes secondary. A constraint framing preserves agency. You steer among admissible options. The Law does not excuse harm or inaction by appealing to fate. Instead, it posits that ordinary choices, efforts, care, and repair are the channels through which balance is achieved. Moral responsibility remains intact. If balance occurs, it occurs through human action, not instead of it.

4.2.7 Predictive specificity you can audit

A constraint hypothesis yields quantitative predictions about where, when, and how specific effects should appear. Under LoF we predict:

- Where: In control and valuation hubs (rIFG, ACC, vmPFC, insula), a residual signal correlated with compensability $\Phi(\hat{L}(t), H_t)$ after controlling for utility, conflict, and arousal.
- When: That residual should strengthen as the horizon H_t shrinks, and compensatory swings in affect (in HCU units) should increase near closure, provided compensatory channels are available.
- How: Dream affect should invert relative to prior-day waking affect beyond what circadian or adaptation models predict; negative days should be followed by systematically more positive dream states if the counterweight mechanism operates.
- Across individuals: In longitudinal cohorts approaching T , partial ledgers $L(t)$ should show variance compression, converging toward neutrality as closure approaches, assuming measurement noise is appropriately modeled.

A miracle-based account does not specify where or when such effects must occur. Constraints predict measurable residual patterns.

4.2.8 Engineering analogies that actually compute

If the Law is real, it should resemble known mechanisms in engineering and control systems: rate limiters, saturation bounds, safety interlocks. These mechanisms do not choose actions; they impose limits on states to prevent catastrophic outcomes. Likewise, the Queue System can be understood as a rising shadow price λ_t on irreparable ledger imbalance as time to fix it shrinks. This maps onto concrete mathematics (see Section 3.6) and aligns with control-theoretic intuitions. In this framing, LoF is a constraint layer atop existing regulatory processes; it can be modeled and quantified without invoking purpose.

4.2.9 Ethical clarity

A teleological framing tempts moral complacency: if the cosmos will make things right, why intervene? The constraint view removes that temptation. Duties remain immediate and local. If someone is in pain, we relieve pain. If someone is isolated, we provide connection. If a relationship is fractured, we attempt repair. If the Law holds, these actions are the channels through which balance is realized. If the Law fails, those actions remain morally required. Either way, responsibility does not shift to fate.

4.2.10 What would force us back to miracles – or away from the Law

What would require abandoning the constraint interpretation? If well-powered studies repeatedly show terminal ledgers outside $\pm K$ under valid measurement; if predicted

signatures (horizon scaling, dream inversion, QS-residuals, variance compression) consistently fail to appear; and if a rival model without a neutrality constraint explains the data equally well or better, then the constraint account collapses. Invoking miracles at that point would not rescue the theory; it would make it unfalsifiable. The appropriate scientific response would be revision or rejection of the Law of Fairness.

Bottom line: Constraints keep physics, biology, and psychology intact. This book does not ask you to accept a comforting story. It asks you to test a strict, audit-ready guardrail. Under LoF, the lifetime ledger $L(T)$ must close within $\pm K$ at T . If LoF belongs in science, it belongs here: as a guardrail you can attempt to break, not as a tale you are asked to believe.

4.2.11 Where we go next:

Having argued for constraints, 4.3 situates the claim within competing accounts of lawhood. That placement clarifies the strength of the claim and how to test it without drifting into metaphysics.

4.3 No Teleology

The Law of Fairness does not say the universe wants or intends balance. It says that histories which would preclude balance are inadmissible for a unified conscious stream. That is a description of a regularity, not a cosmic purpose or intention. In this section, we replace goal-like language with constraint-based language, show how to discuss apparently “goal-directed” patterns without smuggling in hidden intentions, and list the observations that would falsify this strictly non-teleological interpretation of the Law.

Teleological explanations invite two scientific failures:

- Explanatory leakage: Saying “the system aims to restore balance” may sound meaningful, but it merely renames the outcome as a goal. It explains nothing about mechanism. If “aiming” reduces to specific neural circuits weighting options differently, then we must state which weights shift and where they operate.
- Immunity to evidence: If purposes are allowed, negative results can always be reinterpreted (“the aim was hidden” or “the purpose was different”). A teleological claim can contort to fit failure. A constraint claim cannot: either the forbidden trajectories appear or they do not.

For that reason, we adopt a strict rule: no goal predicates. We do not say “the brain wants X” or “life seeks Y.” We state what is bounded, where bounds are applied, and how this would appear in data. By restricting ourselves to observable constraints and measurable outcomes, we keep the explanation within scientific limits.

4.3.1 The non-teleological translation guide

It is easy to slip into purpose-driven phrasing. The discipline is to translate it immediately into constraint language:

- *Teleological phrasing:* “The organism seeks compensation.” *Constraint:* “Admissible options are those that raise the conditional probability of neutral closure, given the current ledger $L(t)$ and horizon H_t .”
- *Teleological:* “The QS guides you to repair.” *Constraint:* “Control weights increase on repair-related options as H_t shrinks, while competing options fall below the admissible threshold.”
- *Teleological:* “Dreams try to counterweight the day.” *Constraint:* “Dream affect shows inverse coupling to prior-day ledger drift ΔL , after controlling for circadian and adaptation baselines.”

Each translation removes intention and replaces it with conditional probability, weighting, and testable coupling. Every statement must be writable as a model specification or preregistered hypothesis. Nothing rests on metaphor.

4.3.2 Constraint without purpose in familiar science

Orderly patterns in nature are routinely explained without invoking aims:

- Noether's theorem and symmetries: Conservation laws arise from symmetries in physical equations, not because nature “wants” to conserve anything. Time-translation symmetry implies energy conservation. By analogy, if experiential dynamics obey time-invariant constraints, a conserved hedonic boundary (neutral closure at T) could emerge mathematically rather than intentionally. Classical particles extremize action; light follows paths minimizing travel time. These look purposeful but result from variational structure, not foresight.
- Thermostats and safety interlocks: A thermostat does not plan to maintain comfort; it enforces thresholds. Homeostatic physiology (Cannon, 1932) regulates variables through feedback loops without intent. Balance emerges from constraint.
- Evolutionary stability: Traits persist because alternatives are eliminated under selection constraints, not because species pursue goals. Apparent purpose arises from admissibility boundaries in fitness landscapes.
- *Dynamic kinetic stability: Self-replicating molecular populations can maintain steady concentrations through rate balance between replication and decay. Stability emerges from kinetics and competition, not from intention. The persistence of the system reflects constraint structure, not planning.*

The Queue System and the Law of Fairness are proposed in this same explanatory family. They describe a constraint layer on trajectories that yields goal-like outcomes (ledger balance) without positing a planner. LoF states that certain endings cannot occur; it does not say the universe causes particular endings to occur.

4.3.1 How teleology sneaks back in – and how we stop it

Teleological drift often re-enters through language. Common leaks and their corrections:

- Language leak: “The brain wants homeostasis.” Fix: “Homeostatic control mechanisms penalize large deviations.”
- Optimization leak: “The policy optimizes fairness.” Fix: “Policies sampled from the admissible set $\mathcal{A}(t)$ preserve high probability of neutral closure.”

- Meaning leak: “That suffering was meant for growth.” Fix: “Following negative ledger drift, we observe increased rate and magnitude of opposite-signed events beyond adaptation baselines, conditional on available channels.”

In each case, the correction replaces intention with mechanism. Every claim must survive translation into model terms. If it cannot be written in statistical or dynamical language, it does not belong in the theory.

4.3.2 Where the regularity lives (without “aims”)

If LoF is real, the regularity is implemented locally as slight weight shifts within known neural systems:

- Valuation circuits (vmPFC/OFC): If an option materially reduces the probability of future compensation (i.e., deepens irrecoverable imbalance in $L(t)$), its subjective valuation decreases; contexts that preserve compensability receive modest upward weighting. The context-to-value mapping shifts as a function of projected ledger impact.
- Control circuits (ACC, rIFG, related networks): Inhibitory thresholds increase for actions that would move the ledger outside the admissible region and decrease for reparative actions, especially as t approaches T and H_t shrinks. It becomes more difficult to execute non-compensable actions and easier to initiate compensatory ones.
- Interoceptive and autonomic systems (insula and related networks): These weight adjustments are experienced somatically—as unease, hesitation, urgency, or relief. A non-admissible path carries a felt cost; a compensatory path carries felt plausibility.

There is no proposed “fairness module.” The constraint operates through ordinary valuation, control, and interoceptive systems. The claim is that small, systematic weight shifts within these circuits, accumulated across time, enforce the boundary condition that $L(T) \in [-K, K]$ at closure.

4.3.3 What looks like purpose from the inside

People often describe experiences in seemingly teleological terms: “I just knew I had to call her,” “It felt wrong to go through with it,” “At last, it felt okay to rest.” These are first-person descriptions of what we interpret as shifts in QS weights. When the admissible set $\mathcal{A}(t)$ contracts or expands as a function of ledger state $L(t)$ and horizon H_t , the shift is felt as a strong “must” or “must not.” That feeling is not a literal voice or cosmic directive; it is the shadow price λ_t signaling urgency within ordinary decision circuits. The

conviction (“I have to do this” or “I can’t do that”) is the subjective experience of constraint, not evidence of external intention.

4.3.4 The evidential burden for a non-teleological law

If the Law of Fairness involves no guiding purpose, then the evidence must reflect that. A non-teleological LoF should appear as graded biases and regularities rather than abrupt, all-or-nothing interventions. We therefore expect to see:

- Gradual, parameterized weight shifts rather than singular interventions. The biasing effect should scale continuously with ledger imbalance and remaining horizon, not flip on at arbitrary moments.
- Ordinary pathways doing the work—sleep, analgesia, social repair, creative absorption—rather than deus ex machina events. Compensatory episodes should resemble intensified versions of known coping and regulatory processes.
- Residual effects, not total overrides. QS influences should appear as measurable residuals after modeling established drivers (utility, threat, reward, arousal), not as wholesale violations of those drivers.

If data instead appeared genuinely miraculous—events with no plausible neural, physiological, or environmental pathway—then the correct scientific response would not be to insert teleology but to withdraw or revise the Law. A scientific theory does not survive by appealing to magic.

4.3.5 What would count against “no teleology”

Several findings would push the Law either into metaphysics or into falsification:

- Causal overreach: Robust evidence of physical law violations occurring in synchrony with ledger needs—brain events without antecedent physical causes that systematically counterbalance imbalance.
- Non-scaling interventions: Compensatory events that arrive with fixed magnitude or timing, independent of ledger imbalance $L(t)$ or horizon H_t . A lack of scaling would resemble scripted design rather than feedback regulation.
- Agent-external orchestration: Repeated cases of improbable external coincidences produce balance without corresponding QS residuals or physiological signatures in the individual.
- Rival model sufficiency: A competing model with no fairness constraint reproduces neutral closures and the full package of predicted signatures (dream inversion λ , horizon scaling γ_1 , QS residual β_4 , variance compression) equally well or better.

Any of these would either move the Law out of science or refute it outright. The commitment here is explicit: if the data require teleology to save the pattern, the Law fails as a scientific hypothesis.

4.3.6 Clarifying agency without purposes

A constraint can narrow choices without authoring them. The individual still deliberates, updates beliefs, and selects among admissible options $\mathcal{A}(t)$. A non-teleological Law of Fairness does not assign credit or blame; it does not author actions. It states only that certain extreme trajectories—those that would render neutral closure impossible—are not admissible. Within the admissible region, choice remains fully human. The Law may rule out catastrophic imbalance, but it does not script virtue or vice.

4.3.7 Communication discipline (for this book and beyond)

To maintain a strictly non-teleological stance, we adhere to four language rules:

- Use constraint verbs: bound, weight, prune, admit, exclude, scale. Avoid intention verbs such as want, seek, aim, try.
- Refer to observables: ledger drift ΔL , horizon H_t , compensability Φ , QS residuals in specified circuits. Avoid abstract spiritual or metaphorical placeholders.
- Frame neutrality as equivalence: neutrality means $L(T) \in [-K, K]$ under preregistered bounds, not mystical perfection.
- Pair every prediction with a fail condition: dream inversion ($\lambda > 0$) must have a null alternative; horizon scaling ($\gamma_1 > 0$) must be testable; QS residuals ($\beta_4 > 0$) must be falsifiable.

This discipline is methodological, not stylistic. It ensures that every claim remains operationalizable and auditable.

4.3.8 Why this matters for ethics

Teleological thinking can distort ethics. If suffering is imagined to serve a higher plan, it risks rationalization. The constraint framing blocks that move. Present suffering demands present care, regardless of any eventual ledger outcome. If balance occurs, it occurs through ordinary acts of relief, repair, and compassion. If balance does not occur, our ethical obligation to reduce suffering remains unchanged. Nothing in this account permits complacency.

4.3.9 Where we go next:

With lawhood situated, 4.4 tightens our language further. We formalize how to keep intention-talk out and mechanism-talk in, so later analysis does not drift into teleology.

4.4 Lawhood: Best System vs. Governing Law

To call something a “law of nature” grants it status beyond “it usually happens.” It implies that the pattern earns a privileged place in our description of reality. Two philosophical traditions articulate this privilege differently. One treats laws as Best System summaries of all facts; the other treats them as Governing constraints that delimit what can occur. Here we examine how the Law of Fairness (LoF) could be understood under each view and why, in practice, we adopt a constraint-flavored Best System interpretation.

4.4.1 Two ways to earn the word “law”

Best System (Humean): Begin with the full mosaic of events—every neural state, every affective report, every terminal ledger value $L(T)$. A law, on this view, is part of the optimal compression of that mosaic: the simplest, strongest theory that captures the largest regularities. It does not push events; it summarizes them. Under this reading, calling LoF a law means that including the principle “for each unified conscious stream, $L(T) \in [-K, K]$ at death of mind” improves explanatory compression more than leaving it out.

Governing Law (Non-Humean): On this view, laws are genuine constraints with modal force. They do not merely describe patterns; they forbid counterfactual possibilities. If LoF were governing in this sense, then life histories ending far outside $\pm K$ would be not just rare but impossible, given the law. As we’ve formulated it – “For each unified conscious stream, the lifetime ledger $L(T)$ closes within $\pm K$ HCU at the death of mind (with specified identity and measurement rules in place)” – the Law of Fairness has a foot in each camp. We stated it like a factual regularity one could test, but we often talk as if it *constrains* what can happen (forbidding certain endings). Let’s unpack each reading.

The formal statement—“For each unified conscious stream, the lifetime ledger $L(T)$ closes within $\pm K$ HCU at death of mind, under predefined identity and measurement rules”—straddles these views. It is stated as an empirical regularity but functions operationally as a constraint on admissible trajectories.

4.4.2 LoF as a Best System regularity

On a Best System reading, the Law of Fairness earns its title if adding one simple axiom to our scientific theory greatly improves our ability to summarize and predict experience across the board. Roughly, imagine adding this single statement to the “axioms” of psychology/biology:

LoF Axiom: For each unified conscious stream (each person’s life), the lifetime affective ledger $L(T)$ closes within $\pm K$ HCU by the death of mind (under the predefined identity rules, metrics, and boundary conditions).

If incorporating that axiom makes our total description of the world simpler and stronger than it would be without it, then LoF qualifies as a law in the Best System sense.

Why might LoF help compress our description of reality?

- It would replace a tangle of local “mood dynamics” clauses with one global lifetime rule. Instead of needing separate mini-theories for each observed rebound or each recovery from trauma, we have one overarching principle.
- It would let us derive observed signatures (dream counterweights, horizon-based shifts, variance compression near death) as natural corollaries of a single assumption, rather than treating each as a weird separate fact. In other words, LoF could “explain” those patterns in one stroke, whereas a rival theory without LoF might need to individually assume or accommodate each of them.
- It could sharpen the definitions of certain edge cases – for instance, clarifying what counts as the end of a conscious stream (“death of mind”) and forcing consistency in how we measure experience (HCl/HCU units) – thus tidying up exceptions that would otherwise clutter the description.

Cost check: Of course, adding any axiom has a cost – it makes the theory more complex. LoF brings extra baggage like defining a “unity index” for consciousness, specifying HCU units, and setting up closure rules. If the world doesn’t consistently obey LoF, those would be pointless complications. So LoF stays in the Best System only if the compression gain (the simplicity and predictive power it adds) outweighs the added complexity across large datasets.

Empirically, this means: if future data show that including LoF’s constraint lets us explain cross-cultural, cross-species affective trajectories with *fewer fudges* and better accuracy than any theory without LoF, then LoF’s “score” as a Best System law goes up. If, on the other hand, we keep finding more exceptions or need lots of caveats to patch the law, then its score drops. The data will tell us if this principle truly pulls its weight as part of the optimal summary of reality.

4.4.3 LoF as a Governing constraint

On a Governing Law interpretation, the Law of Fairness is not just descriptive – it’s literally a boundary condition on the space of possible life paths. It “rules out” those trajectories that would end too high or too low. In this view, the Queue System (QS) we keep talking about is the *mechanism* by which that global rule is locally enforced: the weights in ordinary circuits that prune away low-compensability options and ramp up as t nears are how the law’s constraint manifests in each person.

Strengths of the *governing law* interpretation:

- It directly captures the “modal” flavor of our claim – all along we’ve been saying certain endings *cannot* happen. That’s governing-law language (“must” and “cannot”). If we really believe LoF has no exceptions beyond measurement error, talking of it as a strict constraint aligns with that.
- It naturally explains phenomena like variance compression near death – if there is a global constraint squeezing trajectories, of course you’d see everyone’s ledger values converging as they approach the terminus. It’s a straightforward consequence of a binding rule.
- It treats a single clear counterexample as fatal. On a governing view, if we ever found a well-verified conscious stream that ended way outside the $\pm K$ range (with no extenuating measurement problems), that alone breaks the law. This crisp falsifiability is scientifically appealing: the law could go from “confirmed” to “disproven” with one black swan observation. (Best System, by contrast, might forgive a few anomalies as long as the overall summary is still best.)

Risks to manage for the governing view: it can accidentally slide back into teleology or some unexplained “oomph” behind the law. If we say “LoF *makes* this happen,” one might start wondering “what force makes it so?” To avoid mystique, we have been careful to: (i) explicitly forbid teleological talk (see Section 4.3), (ii) locate every QS effect in *measurable* brain and behavior variables (no unknown forces, just known circuits doing slightly new tricks), and (iii) demand we observe subtle residual signatures rather than any violations of physics or dramatic miracles. In essence, we interpret “cannot happen” in an empirical, defeasible way: “if our measurements are right, we won’t find such-and-such happening.” That keeps the governing interpretation on the rails of science.

4.4.4 A hybrid we can actually test: Constraint-flavored Best System

For day-to-day research, we take a pragmatic hybrid stance. We talk as *if* LoF were a real constraint (because that guides how we design tests), but we evaluate it like a Best-System hypothesis (keeping score with empirical compression and model comparisons). Key features of this stance:

- **Ontology-light:** We allow ourselves to speak in modal terms (“forbidden histories,” “guardrails”) because that language is intuitively useful and leads to clear predictions. However, when it comes time to cash out those predictions, we do so entirely in terms a Humean skeptic could accept: we look for residuals in data, perform equivalence tests, and check variance patterns. We don’t actually

need to assume any spooky ontology – all claims get translated into observable metrics.

- Scorecard discipline: We will be checking, via formal model comparison (information criteria like WAIC/LOO, or minimum description length, etc.), whether including LoF’s principle yields better out-of-sample predictions than leaving it out. This is straight Best-System thinking: does this rule improve our score in compressing/predicting data? If not, it doesn’t deserve law status.
- Modal humility: We freely use the language of “forbidden histories” and “guardrails” because it’s a useful shorthand for designing experiments (it tells us where to look for effects). But we do so with a wink: we acknowledge that this might just be a clever way to compress a ultimately non-modal reality. In other words, we remain open to the idea that LoF’s necessity might be emergent from many mundane facts, not from some higher metaphysical necessity.

This hybrid approach keeps metaphysics from doing the work that data should do, while still preserving the sharp, testable edges of the claim. We get the best of both: we treat LoF as *if* it governs (so we know exactly what would violate it), but judge it by how well it actually helps us predict and explain the data.

4.4.5 How the readings diverge in prediction and practice

To make this concrete, here’s a side-by-side (also including “what we do” in this book):

Issue	Best System emphasis (LoF as summary)	Governing emphasis (LoF as constraint)	Our approach (“What we do”)
Role of exceptions	Tolerate occasional “anomalies” as long as the overall compression is still best.	A single clear counterexample <i>disconfirms</i> the law (assuming no measurement error).	Treat apparent counterexamples as stress tests – investigate them thoroughly; if they hold up under scrutiny, we <i>downgrade</i> or <i>withdraw</i> LoF.
QS-residuals	Consider them useful predictive features that make our model shorter/stronger (but	Interpret them as mechanistic footprints of a real boundary condition in effect.	We pre-register specific brain regions of interest and require finding these residuals after

Issue	Best System emphasis (LoF as summary)	Governing emphasis (LoF as constraint)	Our approach (“What we do”)
	not necessarily “real” forces).		controlling for other factors.
Variance compression near death	Just another pattern the best summary must account for (nice if the theory captures it).	A direct consequence of a global constraint squeezing outcomes near T.	We test both: look for population-level variance shrinkage and individual horizon-interaction effects on variance (both should be present).
Method of evaluation	Compare models by information-theoretic criteria (AIC, BIC, WAIC, MDL, etc.) – pick the one that best balances simplicity and fit.	Use equivalence testing for final ledgers at T (neutrality) and null hypothesis tests for each predicted signature – any significant violation rejects the law.	We do both: report formal equivalence test outcomes (neutrality within $\pm K$, etc.) and model-selection metrics for LoF vs. rivals in each study.
Language used	“LoF is true because it best summarizes the mosaic of facts.”	“LoF is true because it forbids terminal imbalance (and nature obeys that constraint).”	Use constraint verbs and observable terms (see Section 4.3) throughout. We talk of guardrails and forbidden trajectories to guide intuition, but we always tie it back to data.

4.4.6 Rival frameworks and why they fall short as “laws” here

It’s worth noting that several well-known theories could explain *parts* of what we observe about human affect, but none of them (as currently formulated) ensure a lifetime balance without essentially sneaking in an LoF-like assumption. For example:

- Adaptation and opponent-process theories: These are excellent at describing local dynamics (how we habituate to pleasure/pain, how after-effects can create opposite emotions). However, by themselves they don't guarantee that one's *final* cumulative affect ends up zero. One could imagine a world where opponent processes operate but someone still ends up with a huge net deficit simply because life dealt more blows than opponent processes could counter. To actually *guarantee* terminal neutrality, these theories would have to add some global balancing clause – basically recreating LoF within their framework.
- Predictive coding / Free-energy minimization: These are powerful frameworks suggesting the brain strives for stable, minimal-error states. They can explain why we return to homeostatic set-points or why extreme states are unstable. Yet, nothing in pure free-energy theory *by itself* forces the integral of happiness minus suffering over a lifetime to be zero. A brain could minimize surprise and still systematically be sadder (or happier) than neutral lifelong. So again, to enforce an exact balance by end-of-life, one would have to stick in an extra term or constraint – effectively LoF in disguise.
- Reinforcement learning / Homeostatic drive models: These explain a lot about behavior and affect regulation. But on their own, they also do not entail that *each person's ledger closes*. One can easily conceive of an RL agent or a homeostatic regulator that ends life significantly “in the red” or “in the black” – nothing in standard reward learning says “and by the time the episode ends, cumulative reward = 0.” To force that outcome, one must add a global constraint or penalty term – once again, importing LoF's idea into the model.

It's entirely possible (even likely) that these rival processes *will be part* of the eventual explanation under LoF. In fact, LoF might be *realized through* adaptation mechanisms, predictive coding principles, and homeostatic drives working in concert. They aren't mutually exclusive at all – they are candidates for how the guardrails are implemented. But the point is: none of these well-established theories by themselves constitute a law of fairness. They leave open the possibility of permanently unfair outcomes unless one adds something like LoF's terminal clause. So, while we will borrow insights from these frameworks, we still treat LoF as a distinct hypothesis that stands or falls on its own merits.

4.4.7 Decision rule for lawhood status in this book

So, what will it take for us to anoint the Law of Fairness as a true “law” by the end of this project? We set the bar as follows – across multi-site data with pre-registered measurement protocols and boundary criteria, LoF must:

1. Neutral closure in the vast majority of streams: The lifetime ledger $L(T)$ for most well-adjudicated individuals falls within fixed $\pm K$ HCU bounds (with K defined a priori from empirical anchors, e.g. roughly corresponding to ± 0.15 SD on a composite well-being scale). This should hold with high confidence for large samples.
2. Joint presence of signatures: The key predicted signatures (dream counterweights, horizon-based intensity shifts, QS-residual brain signals, and variance compression) are all observed, with quantitatively “tight” effects (confidence intervals well within our pre-specified equivalence margins in each case).
3. Predictive superiority: A model that includes LoF’s constraint achieves strictly better predictive compression of the data (via metrics like MDL, WAIC, LOO) than any rival model without LoF, across multiple tasks and cohorts.

If all three of those legs hold, we consider LoF to have earned provisional “law” status. If any one leg fails persistently (despite increasing data and refinement), we will demote LoF to at best a useful regularity or scrap it. In other words, LoF needs to deliver on outcomes *and predictions* to justify the bold title of a law. We won’t keep it on a pedestal for its philosophical appeal alone.

4.4.8 What would falsify lawhood outright

By now, it should be clear what critical findings would outright falsify LoF in the strong sense (and, in practice, force us to abandon the theory):

- Reliable terminal imbalance: If we *repeatedly* find well-measured individuals whose final lifetime ledger $L(T)$ sits outside the $\pm K$ neutral bound by a significant margin – *and* these individuals had no extraordinary measurement errors or “missing channels” of compensation – that directly violates LoF’s core claim.
- No QS-residual or horizon effect: If rigorous analyses show zero evidence of the predicted QS-residual in control/valuation brain regions, and no sign that people’s choices tilt more compensatory as horizons shrink, then the micro-dynamics part of the theory fails.
- No variance compression: If as people approach end-of-life we do not see any compression of the distribution of net outcomes (for example, variance stays the same or even increases), then the idea of a narrowing trajectory range is falsified.
- Constraint-free model fits as well: If some rival model with *no* fairness constraint can match or exceed all of our results (neutral closures, horizon scaling, etc.) in explanatory power and predictive accuracy, then adding LoF was unnecessary.

Any one of these observations, if confirmed with high-quality data, would be enough to declare that LoF is not a law of nature. We would then either retreat to saying “maybe there’s just a tendency toward balance, not a law” or focus on more local explanations for the phenomena. We want to stress: our bar for calling this a *law* is very high, as it should be. One clear counterexample is all it takes to break it under these criteria.

(In practice, we will treat any serious anomaly as a cue to investigate further – e.g. check if the person truly had a “unified conscious stream” or if measurement invariance held across sites – but the spirit remains: a law that’s repeatedly broken is no law at all.)

4.4.9 Why this section matters for readers across camps

Whether you’re philosophically cautious or adventurous, you don’t have to “buy into” any metaphysics to engage with this theory. If you are metaphysics-averse (the Humean at heart), you can treat LoF as a concise empirical axiom that *might* organize diverse data better than a bunch of disconnected hypotheses. You can remain agnostic about *why* it works and just see if it improves our models. If you’re the type more comfortable with the idea of real laws (the “governing laws” fan), you can view our findings as evidence that there is indeed a real constraint at work in conscious lives, implemented via normal physiology. Either way, our approach is to let shared empirical tests decide the issue, not one’s philosophical leanings. The proof (or disproof) of LoF will come from data – compression metrics, residual analyses, equivalence tests – that everyone can agree on, rather than debates about what a “law” truly is.

Take-home: In this project we behave as *constraint-realists* in our methods and *Best-System minimalists* in our metaphysics. We’ll call LoF a “law” only if it (i) demonstrably compresses/predicts reality better than any rival and (ii) leaves the distinctive, falsifiable footprints of a boundary acting through everyday mechanisms. If it can’t do both, then it hasn’t earned the title.

4.4.10 Where we go next:

Finally, 4.5 lists the practical safeguards—against optional stopping and irregularity—that keep tests clean. These are the habits that prevent us from flattering the theory.

4.5 Research Notes: Optional Stopping and Regularity

These research notes dive into some technical underpinnings. We formalize the “death of mind” as a stopping time in a stochastic process, show how under LoF the affective ledger can be treated as a kind of regulated (super)martingale, clarify when the Optional Stopping Theorem (OST) does or does not apply, and outline how we test the Law’s predictions in practice, including censoring (dropout near death), sequential analyses, and ensuring results are not artifacts of peeking. These notes provide the mathematical backbone for claims about drift regularization, and they describe the statistical safeguards we use to avoid fooling ourselves when testing neutrality at life’s end.

4.5.1 Setup: the ledger as a stochastic process with a stopping time

Let $(\Omega, \mathbb{F}, \{\mathbb{F}_t\}_{t \geq 0}, \mathbb{P})$ be a filtered probability space representing an individual’s unified conscious stream. Here \mathbb{F}_t encodes all information available up to time t (experiences, physiological states, reports, and related observables).

- Instantaneous net affect: $F(t)$ denotes the person’s momentary felt affect at time t , measured via our latent Hedonic Composite Index (HCI) (see Section 3.6). Positive $F(t)$ means net happiness; negative means net suffering at that instant.
- ledger: $L(t) = \int_0^t F(s) ds$ (in Hedonic Currency Units, HCU) is the integrated affective balance from the start of life up to time t .
- Stopping time T (“death of mind”): Define $T = \inf\{t \geq 0 : \text{UnityIndex}(t) < \theta \text{ AND no return within window}\}$, where UnityIndex is an operational measure of whether the person’s conscious processes have irreversibly disintegrated (as defined in Section 3.5), and θ is a threshold (e.g., an isoelectric EEG or equivalent criteria). By construction, T is a stopping time with respect to $\{\mathbb{F}_t\}$.
- Neutral closure (LoF claim): The Law of Fairness asserts that $L(T)$, the final ledger value at the stopping time, lies within $(-K, +K)$ under the preregistered equivalence criterion under the predefined equivalence criterion (i.e., within $\pm K$ HCU, where K is a small positive bound set as “effectively zero”). In statistical terms, it is an equivalence claim: we expect $L(T)$ to be statistically indistinguishable from 0 (neutral) within the preset margin $\pm K$, given the predeclared identity and measurement rules.

4.5.2 Martingale framing (why we talk about it at all)

Write instantaneous affect as $F(t) = \mu(t) + \varepsilon(t)$. Here $\mu(t)$ is the predictable component of affect given the recent past (baseline mood, context value, control costs, learned responses), and $\varepsilon(t)$ is a mean-zero innovation at time t ; in continuous-time form, we model that innovation with a Brownian term $\sigma dW(t)$.

Under the Law of Fairness, we posit that drift $\mu(t)$ is regularized by the current ledger and horizon. Specifically, $\mu(t) = -\lambda \Psi(L(t), H_t)$, where $\lambda \geq 0$ is the strength of the restoring force and $\Psi(L, H)$ is odd in L ($\Psi(-x, H) = -\Psi(x, H)$), increasing in $|L|$, and grows in magnitude as the horizon H_t shrinks.

Define the transformed process

$$M(t) = L(t) + \lambda \int_0^t \Psi(L(s), H_s) ds. \text{ Differentiating, } dM(t) = dL(t) + \lambda \Psi(L(t), H_t) dt, \text{ while} \\ dL(t) = \mu(t) dt + \sigma dW(t) = [-\lambda \Psi(L(t), H_t)] dt + \sigma dW(t). \text{ Hence, } dM(t) = \sigma dW(t)$$

under the standard Brownian-noise formulation with diffusion σ . Under standard SDE regularity conditions (e.g., Lipschitz continuity and integrability ensuring existence of a non-exploding solution), $M(t)$ is a local martingale (and, under suitable integrability, a true martingale) with respect to $\{\mathbb{F}_t\}$.

If $\lambda = 0$, then $\mu(t) \equiv 0$ under this specification and $L(t)$ reduces to a driftless Brownian motion, so nothing enforces terminal closure. If $\lambda > 0$ and Ψ is well behaved, $M(t)$ behaves like a martingale (or like a super/sub-martingale under weaker conditions), enabling Optional Stopping-type reasoning.

4.5.3 Optional Stopping Theorem (OST): when it applies and when it does not

A classical result in probability theory, the Optional Stopping Theorem, gives conditions under which the expected value of a martingale at a stopping time equals its initial expected value. One version states: if $\{M(t)\}$ is a martingale and T is a bounded stopping time (or if suitable integrability conditions such as uniform integrability hold), then

$$E[M(T)] = E[M(0)].$$

In our context, $M(0) = L(0) + \lambda \int_0^0 \Psi(\cdot) ds = L(0)$. So $E[M(0)] = E[L(0)]$, which we may take as 0 if initial ledgers are normalized to zero by definition. Under the stated OST conditions, $E[M(T)] = E[L(0)]$. But recall: $M(T) = L(T) + \lambda \int_0^T \Psi(L(s), H_s) ds$. Therefore,

$$E[L(T)] = E[L(0)] - \lambda E[\int_0^T \Psi(L(s), H_s) ds], \text{ provided the integral term is integrable.}$$

This equation shows what OST actually gives us: the final expected ledger equals the initial expected ledger minus a correction term arising from the compensatory drift. It does not directly give $E[L(T)] = 0$ unless that correction term produces the needed cancellation. Key conditions to check for OST: For the above to hold rigorously, several conditions must be satisfied (we do not simply assume them; we verify them in modeling and diagnostics):

- Integrability / uniform integrability: We need increments of $F(t)$ (and thus $M(t)$) not to be too wild. Practically, $F(t)$ cannot have infinite variance over finite intervals,

and $M(T)$ must have finite expectation. This is plausible for human affect, but must be checked in extreme cases and in heavy-tailed regimes.

- Predictability of Ψ : The drift adjustment $\Psi(L(t), H_t)$ must depend only on information up to time t (no dependence on the future). We define it as a function of the current ledger and horizon estimate, so this holds by construction.
- Regularity of stopping time: T should be almost surely finite (real lives end in finite time) and either bounded or such that $M(t \wedge T)$ is uniformly integrable. Human lifespans are bounded in practice; we can also impose an explicit analytic bound (e.g., 120 years) as a conservative limit.

Now, importantly, what OST does not give us: it does not by itself guarantee $E[L(T)] = 0$. It gives a relationship involving the drift integral. More critically, LoF's claim is per-stream neutrality at closure, not mean neutrality across streams. OST speaks to expectations across an ensemble; LoF speaks to each unified conscious stream.

In summary, OST is a useful guide. Under stated conditions, it constrains $E[L(T)]$ via the compensatory drift term. But LoF goes beyond OST. It asserts neutrality at closure for each stream, not merely that the mean across streams is near zero. We therefore treat OST as a consistency check and a modeling tool, not as "proof" of the Law. The Law stands or falls on empirical tests.

4.5.4 Doob decomposition: LoF as a regularity constraint

Doob's decomposition states that any integrable, adapted process can be uniquely split into a predictable part and a martingale part. For the ledger, we can write:

$$L(t) = A(t) + M(t),$$

where $A(t)$ is a predictable, finite-variation process and $M(t)$ is a martingale term (as in 4.5.2). LoF asserts that, for admissible lives, the predictable component contains a restoring drift term:

$$A(t) = A_0(t) - \lambda \int_0^t \Psi(L(s), H(s)) ds.$$

Here $A_0(t)$ is the baseline predictable trend that would exist without a fairness constraint, and the $\lambda\Psi$ integral represents the compensatory tilt LoF adds. The claim is that by stopping time T , this adjustment confines $L(T)$ to a small interval around 0 under the equivalence criterion.

Two concrete implications that can be tested empirically:

- Horizon scaling: The sensitivity of drift to ledger imbalance, $|\partial A(t)/\partial L(t)|$, should increase as horizon H_t decreases. In other words, closer to closure, the feedback correction per unit imbalance becomes stronger.
- Variance compression: The cross-sectional variance of $L(t)$ (and in particular $L(T)$) across individuals should decrease as $t \rightarrow T$, conditioning on comparable data quality. Early in life, cumulative sums may diverge; near closure, LoF predicts convergence into a tighter range.

These are simplifying theoretical statements; real data include noise, differing horizons, and measurement limitations. The point is that drift regularization leaves signatures such as stronger correction near closure and reduced spread at the endpoint.

4.5.5 Equivalence at a stopping time: testing without bias

Testing LoF's neutrality claim is delicate because everyone's life ends at different times, and the endpoint may correlate with unobserved states. Classical tests can be biased if we "keep checking" until a desired condition is met. We therefore ensure our tests for $L(T) \approx 0$ are not artifacts of peeking or censoring. Safeguards include:

- Preregistration of criteria: Before looking at outcome data, we lock in all key parameters: the neutral equivalence bound K , the unity threshold θ that defines T , any stabilization window after crossing θ (e.g., declaring death of mind only after a prespecified duration of sustained unity loss), and rules for adjudicating ambiguous cases.
- Sequential alpha spending (frequentist): If we plan interim analyses in longitudinal cohorts (e.g., checking aggregate signatures when a prespecified number of endpoints have occurred), we use alpha-spending functions such as O'Brien–Fleming to control Type I error. We do not "peek" at individual neutrality during life; the endpoint is evaluated once at T .
- Bayesian approach: Alternatively (or additionally), we use Bayesian inference with explicit priors over $L(T)$ and report $\Pr(-K < L(T) < K | \text{data})$. Under correct modeling, Bayesian updating remains coherent under sequential inspection, though model misspecification remains a real risk and is addressed via diagnostics.
- Censoring adjustments: Because dropout and missingness often increase near T , we handle informative censoring using inverse probability weighting, joint modeling of outcomes and dropout, and prespecified sensitivity analyses (including worst-case assumptions). If data near T are too sparse or uncertainty too wide, we issue an Inconclusive verdict rather than overstate certainty.

The goal is to avoid classic optional-stopping pitfalls (e.g., “sample until neutral”). We also separate the data used to define T from the data used to test outcomes at T to prevent circularity.

4.5.6 Sequential designs we actually use

We distinguish between cohort-level signatures and individual-level neutrality:

- Cohort-level signature tests: For horizon scaling, dream effects, QS residuals, and related signatures, we can use group-sequential designs. For example, we may plan interim analyses after prespecified numbers of endpoints, using alpha-spending functions to stop early for overwhelming evidence or futility.
- Individual-level neutrality: We do not repeatedly test each individual’s ledger for neutrality during their life. Neutrality is evaluated once, at T . There is no sequential test for an individual’s $L(T)$; there is one endpoint and one verdict. This avoids tautology and the “keep observing until neutral” trap.

In sum, sequential methods apply to secondary markers and cohort aggregates, not to the primary per-person endpoint. The primary endpoint is assessed at T with prespecified criteria; we either observe neutrality within bounds or we do not.

4.5.7 Simulation checks (pre-registered)

Before relying on our analysis pipeline, we validate it with preregistered simulation studies in which the ground truth is known. We simulate scenarios to confirm calibration, false-positive control, and power:

- Null world (no fairness constraint): Generate lives with $\lambda = 0$ (no compensatory drift), with ordinary adaptation and random shocks. We expect no variance compression, no systematic horizon effects, and frequent non-neutrality beyond reasonable $\pm K$. The pipeline should flag this: equivalence tests fail, and model comparisons favor adaptation-only models.
- LoF world (constraint holds): Simulate with drift term $-\lambda\Psi(L, H)$ plus realistic noise and dropout. We expect the full suite of signatures: neutrality within $\pm K$, dream inversions, horizon effects, and shrinking variance near closure. The LoF-based model should outperform adaptation-only models.
- Adversarial worlds: Stress-test heavy-tailed shocks, non-ignorable dropout, and Ψ misspecification. We verify that false-neutral declarations are controlled and that we retain power under noisy, biased, or mis-modeled conditions.

We preregister these simulation plans and publish code, random seeds, and simulated datasets. This provides confidence that any success or failure in real data reflects the world, not pipeline artifacts.

4.5.8 Horizon as a predictable process

A practical challenge is that remaining time H_t is not directly observable. We treat H_t as an $\mathbb{F}_{\{t-\}}$ -predictable covariate: at any time t , we form an estimate H_t based on all information available just before t . This can come from actuarial tables, clinical prognoses, biomarkers, and related predictors (Section 6.4 expands on defining H).

LoF predicts a specific interaction between ledger state and horizon. Schematically:

$$E[F(t) | \mathbb{F}_{\{t-\}}] = \mu_0(t) - \lambda, \text{sign}(L(t)) h_1(|L(t)|) h_2(H_t),$$

where h_1 is a function of ledger magnitude and h_2 increases as H_t decreases (i.e., correction strengthens as remaining time shrinks). This expresses drift regularization: corrective bias grows with imbalance and with a shrinking horizon.

Our testing plan fits hierarchical regression models. For each subject, we include interaction terms of the form $\text{sign}(L(t)) \times h_1(|L(t)|) \times h_2(H_t)$, with random effects per subject to capture baseline $\mu_0(t)$ and sensitivity differences. We use one-sided priors that $\lambda_1 > 0$ and preregister functional forms for h_1 and h_2 (e.g., linear or logarithmic forms justified by theory or prior data).

4.5.9 Where OST can mislead (pitfalls and fixes)

While martingale theory guides reasoning, it can mislead practical inference if conditions fail:

- UI failure: If $M(t)$ (as defined in 4.5.2) lacks uniformly integrable increments, classic OST conclusions can fail. Fix: We use boundedness assumptions or heavy-tail priors that keep expectations finite, and we may saturate Ψ at high $|L|$ to keep drifts finite. We also use robust models (e.g., t-distributed innovations) so extreme events do not dominate inference.
- Optional-peek bias: We disallow repeated individual neutrality checks before T . For cohort interim analyses, we use prespecified alpha spending.
- Censoring near T : If missingness increases near closure and is not modeled, neutrality can be spuriously inflated. Fix: joint missing-data models and sensitivity analyses; if results become ambiguous under plausible missingness assumptions, we conclude Inconclusive rather than overclaim.

- Post hoc redefining T: We forbid adjusting the endpoint after seeing outcomes. Fix: preregistered endpoint rules and public “closure packets” (see Section 3.5.7) documenting UnityIndex trajectories and adjudication decisions.

By anticipating these pitfalls, we aim for fail-safe inference: if LoF passes, it is not because of analytic loopholes; if it fails, we trust the failure.

4.5.10 Distinguishing LoF from adaptation using regularity

A key question is how to distinguish LoF from ordinary adaptation or homeostasis. Short-term dynamics can look similar. A pure adaptation model might posit:

$$dF(t) = -\kappa F(t) dt + \sigma dW(t),$$

a simple Ornstein–Uhlenbeck process. This yields mean reversion in $F(t)$, while the integrated ledger $L(t)$ accumulates noise over time. Such a process can show transient balancing tendencies, but it does not guarantee compression at end of life; under adaptation alone, some individuals can end with large net positives or negatives by chance.

LoF’s addition of $\Psi(L, H)$ changes two things:

1. Reversion strengthens as H_t shrinks (horizon effect).
2. Drift depends on $L(t)$ itself, not only on $F(t)$. The system “tracks the integral,” not just the instantaneous deviation.

Differences we test:

- Ledger-variance slope near closure: Under LoF, variance of $L(t)$ across individuals should flatten or decline as cohorts approach closure. Under adaptation alone, variance tends to continue increasing or plateau at a higher level.
- Predictive gain from the $\Psi(L, H)$ term: We compare predictive models with versus without $\Psi(L, H)$. If LoF is real, including $\Psi(L, H)$ should improve forecast accuracy or likelihood under cross-validation.
- Counterweight timing (dreams and related processes): If dreams provide counterweights, dream affect should correlate with prior-day ledger drift in a compensatory direction beyond circadian effects and baseline tendencies. Adaptation alone might yield rebound, but LoF predicts proportionality to the size of drift and specificity tied to the ledger state.

Any single discriminator may be inconclusive, but together they form a constellation. If we observe variance compression, improved prediction from $\Psi(L, H)$, and timed counterweights, then adaptation alone is insufficient.

4.5.11 Minimal checklist for lawful optional stopping

Bringing this together, when testing LoF at a stopping time with equivalence, we adhere to:

- Fixed endpoint definition: Stopping time T (death of mind) is defined by objective criteria (UnityIndex threshold θ plus a prespecified no-recovery interval), preregistered and not altered after outcomes are known.
- Integrability constraints: We verify that $F(t)$ has finite variance increments over relevant intervals and that $\Psi(L, H)$ is bounded or grows slower than linear where required for OST-style reasoning.
- Predictability of inputs: Any covariate conditioned upon $(L(t), H_t)$ is estimated using information available up to $t-$, not future data.
- No interim individual tests: We do not test neutrality for individuals prior to T . For cohort interim analyses, we use proper alpha spending.
- Equivalence testing at T : At final analysis, we use TOST or Bayesian intervals to assess whether $L(T)$ lies within $\pm K$, and we apply prespecified criteria for related signatures (e.g., variance compression margins).
- Transparency: We release closure reports, data, and code used in final analyses subject to privacy protections.

This checklist ensures that if we declare support for LoF, it is earned, and if LoF fails, it fails under fair and auditable conditions.

4.5.12 Take-home

Optional stopping is not a proof machine. It is a warning label: endpoint-based testing is fragile unless rules are prespecified and missingness is handled honestly. We frame LoF's neutrality as an equivalence test at stopping time T and state the regularity conditions under which those tests are valid. In plain terms, to test LoF fairly we formalize the ledger process and impose ground rules: no peeking, prespecified endpoints, and explicit missing-data handling.

The Law of Fairness itself goes beyond optional stopping mathematics. It asserts a stronger outcome: not merely that the mean ledger across lives is near zero, but that each life ends within a narrow neutral band by closure. It claims this occurs via predictable, horizon-weighted drift, a drift we can model, measure, and attempt to disconfirm using the methods above.

In the chapters to come, we implement these methods and see whether data uphold the Law. We identify the weak points (where it could fail) and the signatures (where it should stand out), then test accordingly.

4.5.13 Where we go next:

Part III turns from framing to mechanism. We move into concrete tasks, field data, and adversarial designs that let the claim either survive pressure or fail in daylight. We shift from big-picture lawhood to the on-the-ground mechanism that could implement it. Chapter 5 explores the Queue System (QS) – our model of how the mind might locally enforce the fairness constraint. We'll see how QS places a “shadow price” λ_t on future suffering as the horizon shrinks, biasing which thoughts and actions come “next” in line. This takes us into the realm of cognitive control and neural substrates, bridging the abstract law with day-to-day human psychology and biology.

Part III — How the System Works (From the Inside)

Life is lived from the inside out. Think of the last time you felt an inexplicable tug in your chest to reach out and help someone, or a gut feeling that gently steered you away from a bad decision. Where do those subtle impulses come from? In this Part, we turn to how fairness might actually work itself out within a life, not as a mystical force, but as the quiet reweighting of choices you experience every day. Parts I and II set out the Law of Fairness and defined its boundaries; Part III now moves inside experience, to the domain of attention, urges, and options as they appear or fade moment by moment. We won't invoke metaphysics or cosmic intention. Instead, we'll see how a global requirement (that each life's ledger closes neutral at the end) could be implemented through ordinary psychological and neural machinery. We ask: what would it feel like from the first-person perspective if such a fairness law is in operation, and what footprints would it leave for scientists to measure?

By "inside," we mean the person's own point of view, along with the immediate brain and body processes that shape that view. This is not a mystical realm. It's the everyday way your mind manages choices. When you feel a spontaneous tug to call someone, a soft brake on a rash remark, or a sudden sense of ease after long strain, something in your control system is quietly shifting. Part III gives names to these subtle shifts, pinpoints where they occur in the brain, and explains why they grow stronger as one's remaining time horizon shrinks. Importantly, nothing here requires a new "fairness module" or a ghost in the machine; it's all known cognitive processes doing a new job under a strict constraint. In short, we explore mechanisms that could enforce the Law from within, if that Law is real.

The organizing concept in this Part is what we call the Queue System (QS). This is our model for the Law of Fairness acting locally in the mind. QS is not a little person in your head or a grand planner making moral judgments. It's more like an internal scheduling layer that silently shapes what options you consider and how strongly each option calls for your attention. QS doesn't override your free will or insert new desires; you still make your choices. In decision-theoretic terms, this is equivalent to introducing a dynamic Lagrange multiplier on the fairness constraint: QS adjusts option values to favor trajectories that keep the eventual ledger balanced. Practically, this means QS raises a "shadow price" on any future that isn't compensable given the current situation. As your cumulative ledger $L(t)$ drifts far from zero, or as time grows short, certain paths become increasingly "expensive" internally. From the inside, this isn't experienced as an explicit price tag, of course; it's felt as intuitive somatic markers nudging your decisions (a felt "no, not that," a felt "yes, this is right," or a felt "wait, hold off" at crucial moments). Later

in this Part we'll formalize these ideas, but here's a preview: if we let $\hat{L}(t) = \int_0^t HCl(\tau) d\tau$ be the running total of the Hedonic Composite Index (HCl, our measure of net momentary affect), and let H_t be the “effective time remaining” at time t , we can define a compensability function $\Phi(\hat{L}(t), H_t)$ that gauges how much a given action would improve the odds of finishing with $L(T) = 0$ at the end. The Queue System hypothesis is that certain brain circuits add a term $\beta\Phi(\hat{L}(t), H_t)$ into their decision-making signals, effectively shrinking the set of viable actions to those that keep the ledger balanced. In a simple model, we'd expect to see something like:

$$A(t) = f(\text{utility, conflict, arousal}) + \beta\Phi(\hat{L}(t), H_t) + \varepsilon,$$

with $\beta > 0$ increasing as H_t decreases (i.e., as the horizon closes in). In plainer terms, the closer you are to life's endpoint (or any point of no return), the more heavily your mind should weight actions that restore balance, and the more it should inhibit actions that would leave no time to recover. We'll see concrete predictions of this effect, and how to test for it, throughout Part III.

Before diving into data and math, let's ground this in everyday experience. We will track three broad channels of experience that QS would influence:

- Valuation shifts – The brain's reward circuits (vmPFC/OFC) quietly change the “flavor” of options as your ledger and horizon change. Contexts or actions become more attractive if they help correct an imbalance, or more aversive if they would deepen it. (Think of how an option that once seemed tempting later feels pointless or unappealing when you're already hurting or running out of time; the taste of that option has changed.)
- Control thresholds – The brain's internal “go/no-go” signals (in ACC, rIFG, and basal ganglia) adjust their thresholds. It becomes easier to refrain from an action that would push your experience further out of balance, and easier to commit to an action that would help bring things back toward balance. In plain terms, some impulses get gently braked and some intentions get a green light more readily, based on whether they keep eventual neutrality within reach.
- Interoceptive tone – Bodily feedback (insula, brainstem, autonomic systems) mirrors the changing set of admissible paths. You might feel an unexpected urge to rest, a sudden appetite for something comforting or reparative, an openness to help, or even a new spark of creativity at odd times, all subtle bodily nudges aligning you with actions that heal or stabilize your ledger. Crucially, these nudges carry a direction without compulsion. They don't force you, but they gently bias you toward choices that “fit” the fairness constraint.

(Each of these channels, on their own, operates in perfectly ordinary ways in any person. What's different under the Law of Fairness is the residual pattern they collectively produce. After we account for all the usual factors (basic utility, conflict monitoring, stress arousal, etc.), there should be something left over: a distinctive tilt in these signals that correlates with $\Phi(\hat{L}(t), H_t)$. That residual will be our empirical handle on QS.)

If the Law of Fairness truly holds, what should it feel like to live under this law? In a word: nothing supernatural, just a series of gentle course corrections that might otherwise go unnoticed. You would experience nudges, not commands, in your stream of thought. Some ideas or urges would arrive with more weight than others, for reasons you can't quite articulate; some temptations would fade on their own, while certain intuitions (to reach out, to apologize, to take a rest) would stick around despite distractions. As time goes on, especially if your remaining time becomes short, you'd likely notice a kind of time-weighted clarity about what matters. For example, people with a serious illness or in advanced age often report that trivial or "not worth it" activities lose their appeal, whereas acts that bring peace, closure, or comfort feel urgently right. It's as if the mind quietly knows there's no time for tangents. You might also notice low-drama counterweights in your life: after a period of intense pain or stress, you get an unusually deep sleep or a disproportionate sense of relief from a small kindness. A small positive experience lands much bigger than usual, as if to counterbalance recent negativity. None of these effects announce themselves as cosmic or magical; they just accumulate into a life that, in retrospect, finds its way back to balance. In sum, the inside story of an enforced fairness law would be one of subtle tilts that help a life reach a neutral ledger by the end.

Now, what should outside observers expect to see if the Law of Fairness is operating? Part III lays out several testable signatures. First, we predict detectable QS-residual signals in the brain's control and valuation centers (rlFG, ACC, vmPFC, insula) that correspond to our compensability metric Φ . In practice, this means if scientists measure neural activity during decision-making and statistically factor out everything ordinary (reward value, conflict level, arousal, etc.), there should remain a small but systematic signal that tracks how much a choice would improve the chances of closing the ledger neutrally. Second, we expect to see horizon scaling effects: the closer someone perceives themselves to the end of the line (the smaller H gets), the stronger the brain's inhibitory brakes for irreparable actions should become, and the more pronounced the bias toward reparative or balance-preserving actions. In behavior, the person's choices should shift accordingly as well. Third, we anticipate counterweight dynamics during sleep: dream moods should inversely mirror recent waking imbalances (a particularly hard day might be followed by a soothing dream, as if the mind provides overnight relief

when waking life didn't). And fourth, as lives approach their endpoints, we expect variance compression across individuals' outcomes. In plainer terms, as different people near the end of life (and assuming they have access to care and support), their cumulative felt-experience scores should converge toward neutral within a narrow margin. By our definitions, a near-neutral closure would mean a final ledger within ± 0.15 standard deviations of zero, with a final drift slope within ± 0.05 SD/day of zero in the last stretch. We also predict the spread of outcomes between people will shrink; the final ledger variance should be no more than about 80% of what it was mid-life, given quality data. In essence, individual life trajectories that might have been very unequal in mid-course will show more similar outcomes at the finish line, provided the Law is at work. Each of these patterns—neural residuals, horizon effects, counterbalancing dreams, and end-of-life convergence—is concrete and falsifiable. In Chapters 5 and 6, and later in Part IV, we will define exactly how to measure these effects (with preregistered analyses and thresholds) so that rival theories can be fairly tested. If those predicted patterns don't show up, the Law of Fairness fails.

Throughout our explanations in Part III, we enforce three strict guardrails to avoid any teleological or moralistic misreading of the theory. First, no purposeful-universe language: We never say the brain "wants" balance or that nature "seeks" fairness. We instead describe the dynamics as the inevitable pruning of inadmissible trajectories; the system follows a constraint, not a conscious goal. Second, no moral smoothing: We will never imply that future balance justifies present pain. If the Law operates, it does so through timely relief and support; suffering is always a call for care, not something to be rationalized away. Third, no heroics or melodrama: The effects we're looking for are mundane and compassionate in nature: quality sleep, effective pain relief, reconciliations, small acts of help or creativity. We are not searching for grand cosmic justice events. The point is that if fairness is conserved, it achieves that through ordinary means and human interventions, not through any miraculous deus ex machina.

What this Part will do for you:

- Translates the lofty fairness claim into on-the-ground mechanisms and predictions.
- Introduce the Queue System (QS) as a one-sentence concept and a detailed model, showing how a fairness constraint could be realized through everyday mental processes (no spooky forces required).
- Explain how choices are weighted and pruned: You'll learn how the mind's "menu" of possible actions gets filtered down by QS to keep a life's trajectory compensable. We formalize the idea of choice sets and admissible policies, and

we develop the math of a shadow price on futures that one might not have time to offset.

- Identify measurable signatures of fairness in action: We point out which brain regions and behavioral patterns to monitor for QS effects, including the QS-residual signal after accounting for ordinary decision factors, the scaling of this effect as one's time horizon H shrinks, and subtle counterbalancing signs like dream mood shifts. Each signature comes with a way to test it (e.g., regression residuals, time-series analyses) and a criterion for success or failure.
- Reinforce an ethical, human-centered approach: We emphasize throughout that present comfort and dignity are paramount. The Law of Fairness, if real, works through human care, not around it. No argument for eventual balance will ever be used to excuse suffering in the moment. “Relief is a systems variable; comfort and dignity override data collection.” In practical terms, that means any experiment or intervention inspired by this theory must put compassionate care first and knowledge second.

Chapters in this Part:

- **Chapter 5 — The Queue System (QS)** - Introduces QS in one sentence and explains how it weights options rather than picking them. We formalize choice sets and admissible policies, develop the math of the shadow price on incompensable futures, identify neural “seats” of QS (rIFG, ACC, vmPFC, insula), and give everyday examples, all without invoking any spooky forces.
- **Chapter 6 Time Horizons and the Shadow Price** - Shows how expected time remaining (short vs. long horizons) enters subjective life. We explain why nearing the end of the line intensifies QS’s effects (the shadow price skyrockets), with clinical and everyday signatures: think palliative reconciliations, end-of-term clarity, or last-week bursts of creativity when a deadline looms.

Where we go next:

The next chapter turns the law from an abstract claim into a working mechanism. We move from “what must be true in the ledger” to “how the system could make it true from the inside,” introducing the Queue System (QS) as a constraint that trims what can line up next. Chapter 5 begins with the one-sentence account and then builds the pieces we can actually test.

Chapter 5 — The Queue System (QS)

This chapter moves from the life-level claim to the machinery that makes it felt. Imagine your mind as a busy nightclub on a Friday night. There is a bouncer at the door and a line of people waiting to get in. Inside the club is your attention and action space, the dance floor of thoughts and behaviors you might engage in. The bouncer is not telling you whom to dance with or what to do once inside; instead, they control who gets in and how many can crowd the floor at once. This is the essence of the Queue System. It is a metaphor, but a useful one. At any given moment, your mind is teeming with potential thoughts, urges, and impulses jostling for entry into your conscious spotlight. The Queue System (QS) functions like that bouncer and line manager combined. It does not make your choices for you; it manages which choices appear viable and compelling in the first place.

The Queue System is our proposed answer to a central question: if every life must end with a neutral ledger of felt experience under the Law of Fairness, how does an individual life get steered toward that balance from the inside? QS is the machinery we hypothesize could perform that steering quietly, automatically, and through known cognitive processes. QS is not a mystical force or a moral compass; it is the Law's constraint realized through ordinary biology and psychology. You can think of QS as a weighting layer in the mind. At every moment, you have a set of possible actions or thoughts you could entertain. QS works behind the scenes to tilt the odds. It makes some options more salient and easier to embrace, and others less appealing or harder to execute, based on one principle: keep the life on a path that can be brought to neutral by the end. If an option would worsen your ledger in a way that might not be repairable within the remaining horizon, QS reduces its salience or attaches a felt hesitation. By contrast, if an action would help heal, offset harm, or preserve compensability, QS increases its pull. QS governs who gets through the door and how long they remain in your attention, but not what you ultimately do with them. Within the options that line up, you retain agency. QS curates the line so that the options before you remain admissible in the sense of not derailing neutral closure at T.

Let us make this more concrete. On a day when you have recently experienced more pain than relief, and your ledger has drifted negative, QS tends to queue reparative actions closer to the front. You might spontaneously consider calling a friend, resting, or seeking comfort, even if those would not normally top your list. Conversely, if your recent days have been unusually positive and your ledger has drifted upward, QS increases the weight on restraint. You may unexpectedly hesitate before indulging a risk you would normally take. Now add the effect of time horizon H. As remaining time to compensate

shrinks, QS tightens the admissibility filter for actions that cannot be offset in the available time. Options that would create long-term imbalance lose traction. Actions that preserve flexibility or enable repair gain priority. From the inside, this feels like a soft “no” to some options that once seemed fine and a surprising “yes” to rest, help, or meaningful effort. It produces clarity about what matters as H declines. This occurs without explicit ledger calculation. It is an automatic biasing of attention and motivation that feels like intuition.

Where might the Queue System be implemented in the brain? Evidence points to networks already involved in valuation and control. Valuation circuits in the ventromedial prefrontal cortex (vmPFC/OFC) likely incorporate compensability into value signals. Options are valued not only for immediate reward but for their contribution to maintaining $L(t)$ within recoverable bounds. The subjective “taste” of a choice shifts accordingly. Control circuits in the anterior cingulate cortex and right inferior frontal gyrus, together with basal ganglia gating pathways, adjust thresholds. It becomes easier to veto actions that would produce non-compensable imbalance and easier to initiate actions that preserve compensability. Interoceptive systems, including the insula and brainstem autonomic circuits, translate these shifts into felt signals. Heaviness in the gut or a wave of calm can function as somatic markers that tune choice. These systems do not introduce new goals; they modulate the ease of pursuing existing ones so that trajectories remain compensable.

If QS is real, it should leave measurable signatures. After accounting for expected utility, conflict, risk, arousal, and other standard predictors, there should remain a residual component correlated with $\Phi(\hat{L}(t), H)$. We call this the QS-residual. In neural data, we would expect activity in regions such as ACC, rIFG, vmPFC, and insula to show residual variance explained by Φ beyond ordinary decision variables. Behaviorally, we would expect systematic biases toward reparative actions following negative ledger drift. We also predict horizon scaling, with QS-residual influence increasing as H decreases. Sleep provides an additional test case. Dream affect should exhibit inverse coupling to recent waking imbalance beyond circadian and adaptation effects. These predictions are concrete and falsifiable. If such patterns do not appear under controlled measurement, the QS account fails.

One clarification is essential. The Queue System is a constraint-based filter, not a conscious purpose or moral arbiter. It does not command goodness and does not guarantee comfort at any particular moment. It narrows or reweights the set of admissible actions only when necessary to preserve the possibility of neutral closure. Within that admissible set, agency remains intact. If individuals reliably end life with large

uncompensated ledger values under valid measurement and intact channels, the Law of Fairness fails. There is no appeal to hidden intention.

Throughout this chapter, we maintain strict methodological discipline. We use constraint language rather than goal language. We define equivalence bands $\pm K$ in HCU for neutrality. We commit to null hypotheses and falsifiers. QS stands or falls on observable evidence.

What you'll get from this Chapter:

- A summary you can carry into any conversation about how fairness might be enforced in a life.
- An explanation of choice sets and admissible policies—clarity on how, at any given time t , the mind’s set of thinkable actions $\mathcal{U}(t)$ gets narrowed to an admissible subset $\mathcal{A}(t)$ under QS. You’ll see exactly how QS prunes or reweights your options behind the scenes, and we’ll formally define the compensability factor $\Phi(\hat{L}(t), H_t)$ that determines which paths remain open.
- The “shadow price” formalism linking ledger and horizon to felt options—a concrete mathematical model showing how the pressure to balance (quantified by a shadow price λ_t) increases as the ledger $L(t)$ drifts and remaining time H shrinks. In short, you’ll learn how to compute the tilt QS introduces, and why it ramps up as one’s horizon closes.
- A tour of QS’s neural seats and how to detect them – identification of the brain circuits likely implementing QS (including vmPFC/OFC for valuation, ACC and rIFG for control, and the insula for interoceptive tuning). We outline how to measure the QS-residual signal in these areas – the bit of neural activity we expect to find after accounting for standard decision-making factors.
- Relatable examples of QS in everyday life – a collection of low-drama, intuitive examples illustrating how QS might feel from the inside. These include those “nudges”, “brakes”, and surprising moments of “ease” or “clarity” that people often experience, all explained without invoking any mystical or moral forces.
- Clear fail conditions for the QS account – we enumerate what evidence would disprove or seriously weaken QS. This includes specific experimental outcomes (or lack thereof) that would indicate the Queue System model is incorrect, ensuring that the theory is fully testable and not a hand-wavy explanation we cling to regardless of data.

Subsections in this Chapter:

- **5.1 QS in a Sentence** – The compact definition of the Queue System, plus a one-paragraph plain-English translation to illustrate its meaning. We distill the entire concept into its essence so you have a baseline before diving deeper.
- **5.2 Choice Sets and Admissible Policies** – From all thinkable actions $\mathcal{U}(t)$ to the admissible subset $\mathcal{A}(t)$. This section shows formally how QS “shrinks” or reweights the menu of options. We introduce the mathematical definition of Φ (the compensability feature) and demonstrate how the current ledger $L(t)$ and horizon H_t determine which actions remain viable.
- **5.3 Neural Correlates: rIFG, ACC, vmPFC, Insula** – How the Queue System might be implemented in the brain. We detail the expected signatures in specific neural circuits for valuation (vmPFC/OFC), control (ACC and rIFG, including the subthalamic nucleus (STN) for stopping), and interoception (insula). Here we formally define the QS-residual and discuss experimental tasks and measurement plans to isolate it, along with confounds to rule out (to ensure we’re not mistaking ordinary processes for QS).
- **5.4 Dreams as Low-Cost Counterweights** – Why sleep and dreaming might be an ideal workshop for QS to do its balancing. We explain the hypothesis that dreams can tilt one’s affect back toward baseline without expending waking resources. Specific patterns to look for are proposed (e.g. dream affect inversely tracking the previous day’s ledger drift), along with ways to test this “nighttime counterweight” idea through dream content analysis and physiology. Predictions and potential fail conditions for the dream effect are laid out.
- **5.5 Research Notes: QS-Residuals After Nuisance Modeling** – A methodological guide for identifying QS’s unique contribution in data. We outline the statistical approach: first build a strong baseline model of behavior and neural activity using known factors (reward, risk, conflict, etc.), then add in QS-specific features (Φ , horizon terms, shared-resource penalty) and see if they significantly improve predictive power out-of-sample (e.g. via cross-validation or an information criterion like WAIC). We discuss model selection, checks for overdispersion in any count data (start with Poisson; if variance/mean > 1.2, switch to a Negative Binomial with log link), and other pitfalls to avoid. The goal is to ensure any claimed QS effect is robust and not an artifact of mis-specified models.
- **5.6 What Would Falsify QS?** – Concrete, preregistered falsifiers for the Queue System mechanism and the Law of Fairness as operationalized so far. We list specific experiments or observed patterns that would decisively refute the QS

theory (or force us to downgrade the Law of Fairness to a mere tendency). Examples include finding no QS-residual where it should exist, seeing people consistently end life with uncompensated extremes, or alternative models explaining all effects. We even set a decision rule for when to abandon or revise the QS account if enough “strikes” accumulate.

Where we go next:

We begin with the most compact statement of QS—what it does and what it does not do—so the rest of the chapter has a clear anchor. With that sentence in hand, 5.1 sets the tone and vocabulary we will use for the mechanism before we widen to choice sets, neural correlates, dreams, and falsifiers.

5.1 Queue System in a Sentence

The Queue System (QS) is a horizon-sensitive constraint-weighting layer that prunes and reweights the thoughts and actions available to a unified conscious stream so that, given its running ledger $L(t)$ and remaining horizon H_t , only trajectories that preserve a high probability of neutral closure within $\pm K$ at death of mind remain realistically thinkable, selectable, and sustainable.

5.1.1 Plain-language gloss

QS is the quiet policy at the door of your mind's nightclub. It does not tell you whom to dance with; it decides who gets in, who gets waved through, and who never makes it onto the floor. The sorting happens along two contextual variables: how your running ledger $L(t)$ is currently trending and how much effective time H_t you likely have remaining.

When your recent ledger has drifted negative (that is, you've accumulated more pain than relief), ideas that could help repair or offset that imbalance rise more easily into awareness. When your time horizon is short, options that would create irreversible losses or long detours quietly lose traction and fail to hold attention.

Importantly, you still choose among what appears. QS does not pick for you. It shapes the admissible set of options — the ones that remain realistically thinkable, emotionally sustainable, and compatible with neutral closure within $\pm K$ at the end of mind — but within that set, the steering is yours.

5.1.2 What QS changes (and what it does not)

QS changes:

- Availability of options. Some potential actions or urges simply won't occur to you with enough force to pursue, as if they've been deprioritized or filtered out.
- Stickiness of options. Some actions, even if started, cannot be sustained; you lose the thread or motivation, while others stick around and keep engaging you.
- Felt weight of options. Certain choices carry an immediate gut feeling of yes or no (a somatic marker) that skews you toward or away from them, even if you can't rationalize why in the moment.

QS does *not* change:

- Physical possibilities or facts. QS doesn't alter what's objectively possible. It won't give you new abilities, erase memories, or change external constraints; it works within reality.

- Your capacity to choose among remaining options. It never violates free will or agency. Among the options that are presented to your mind with sufficient salience, you still exercise your judgment, preferences, and moral responsibility. QS won't force a particular choice; it shapes which choices feel viable.
- Ethical responsibility. You remain responsible for choices you make within the admissible set. QS not removing a bad option doesn't absolve wrongdoing, and QS narrowing your options doesn't negate praise or blame for how you navigate within those guardrails. In short, QS adjusts the menu, but you order the meal.

5.1.3 Minimal formal hook

Let $L(t)$ be the cumulative ledger at time t (aggregate net suffering/pleasure up to t) and let H_t be a horizon proxy (a measure of how much time likely remains, in an absolute sense or as a proportion of expected life). For any candidate action u available at time t , define a feasibility-of-compensation score: $\Phi(u; L(t), H_t) \equiv (\text{predicted change in } \Pr\{L(T) \in [-K, K]\} \text{ if } u \text{ is pursued})$, i.e., how much taking action u is predicted to raise the probability that by end of life the ledger falls within the neutral band $[-K, K]$. Intuitively, Φ measures how much this option would help (or harm) the chances of ending in balance.

Given this, QS imposes selection weights $\omega(u; t)$ that increase with Φ and effectively shrinks the admissible set $\mathcal{A}(t)$ to those options that keep neutrality feasible. Formally: $\mathcal{A}(t) = \{ u \in \mathcal{S}(t) : \Pr\{ L(T) \in [-K, K] | u, L(t), H_t \} \geq 1 - \varepsilon \}$, where $\mathcal{S}(t)$ is the set of selectable actions given ordinary decision factors (utility, etc.), and ε is a small tolerance (preregistered). All u in the admissible set are then weighted by $\omega(u; t) \propto \exp[\beta \cdot \Phi(u; L(t), H_t)]$, with $\beta > 0$ increasing as H_t shortens (β is the shadow-price gain that intensifies compensatory weighting when time is scarce).

In plainer terms: QS filters $\mathcal{S}(t)$ down to $\mathcal{A}(t)$ based on whether choosing u keeps the odds of eventual neutrality above a high threshold (near 1). If an action would likely make neutrality unreachable (very negative Φ), it gets dropped or severely down-weighted. If an action would significantly improve the chances (positive Φ), it stays in play, and may feel more compelling.

Neurally, QS should appear as an additive residual influence in the usual decision circuits, proportional to this pooled Φ feature (see Ch. 3.6 and 4.1–4.5 for background). In upcoming sections we will look for that residual.

5.1.4 Inside-view feel

What does all this feel like in everyday life? Here are a few recognizable examples of QS at work from the inside:

- *A text you nearly send fades before you hit “send.”* Yesterday you might have impulsively sent that snarky reply, but tonight something in you lifts your finger off the trigger; the thought loses momentum on its own. (In QS terms: the option’s Φ was negative, it would likely cause regret or harm with little time to repair, so it quietly slid out of the admissible set.)
- *An apology that felt impossible yesterday now feels obvious.* You wake up with clarity that you should reconcile or reach out, even though previously pride or fear stopped you. (Perhaps your ledger drifted further negative, making the reparative action’s Φ more positive; QS let it into focus and lowered the internal barriers to doing it.)
- *A nap, a walk, or a song feels disproportionately “right” after a hard week.* Following a period of strain, restorative or meaningful small actions suddenly carry a big yes feeling, as if your body and mind are suggesting a balancing move. (This is QS biasing you toward relief and recovery options when your compensatory need is high.)

These are tilts, not commands, gentle guardrails, not direct steering. You still have to choose to follow through, but QS tilts the field so that certain choices come to you and stick with you more than others.

5.1.5 Outside-view signatures

From a third-person (scientific) perspective, if QS is operating we should be able to measure its effects in the following ways:

- QS-residuals in rIFG/ACC/vmPFC/insula: After accounting for known influences on these brain regions (utility value, conflict/difficulty, arousal/stress, etc.), there should be a residual signal correlated with our Φ (feasibility-of-compensation) metric. Essentially, each region should show activity patterns that cannot be explained by ordinary factors alone, indicating an extra signal consistent with QS.
- Horizon scaling: The shorter the horizon H_t , the stronger these QS-residual signals should become. For instance, as an experiment imposes a sense of impending ending (say a deadline or a prognosis), we expect to see larger residuals and bigger opposite-signed affect swings (stronger biases toward relief or restraint), with QS increasing its influence when time is short.
- Dream counterweights: During sleep, especially REM, we expect dream affect inversely tracking prior-day drift, beyond what general stress or adaptation accounts for. If one’s day pushed the ledger down, dreams that night should disproportionately feature positive or compensatory themes (and vice versa for unusually positive days), more so than baseline homeostatic dreaming would. In

other words, dreams should act as contrastive counterweights to waking imbalance, and we aim to see this with at least moderate reliability (we target $\kappa \geq 0.60$ in coding agreement for dream-content categories like “relief” or “mastery”).

- Ledger variance compression near closure: As individuals approach end of life (or any context of finality), if QS has done its job, their partial ledgers (the accumulated net suffering at that point) should show less variance across people than you’d expect by chance. In plainer terms, people’s lives converge toward neutrality. We should be able to quantify this compression of variance in well-measured cohorts (with the caveat that data quality and measurement invariance must be high for comparisons across individuals).

5.1.6 Fail conditions specific to QS

We also delineate up front what evidence would indicate that QS is not operating as claimed. The following are fail conditions that specifically undermine the QS mechanism (these mirror the formal falsifiers in 5.6):

- Residuals vanish under rigorous controls. If we model brain and behavior with all the usual factors and no extra unexplained signal remains (i.e. adding Φ or horizon terms doesn’t improve predictions at all), then QS has no unique footprint.
- No horizon interaction where there should be. If, in situations where compensatory options exist, shortening the time horizon does not cause any narrowing or tilting of choices (no intensification of brakes or boosts), then QS isn’t doing what we expect in the endgame.
- Admissible-set leakage. If we repeatedly observe clearly non-compensable trajectories proceeding unimpeded near the end of life in well-tracked cases, for example, someone with little time left and available help still spirals further down with zero felt resistance or substitution, that means the supposed guardrails failed exactly when they were needed most.
- Rival accounts match all signatures. If a model with no fairness constraint (say, just homeostatic adaptation plus risk aversion plus fatigue) can reproduce all the phenomena we ascribe to QS (the horizon effects, sleep inversions, etc.) with equal or better predictive power, then QS isn’t a necessary explanation.

Any of the above, if reliably demonstrated, would force us to weaken or abandon the QS account of the Law of Fairness.

5.1.7 One-sentence takeaway you can quote

“The Queue System is how a global fairness constraint shows up locally, by quietly shaping which options are thinkable and keepable so that a life can end in the neutral

zone without anyone breaking the rules of physics or of choice.” (In conversation, you might simply say: “It’s like your mind’s internal bouncer making sure you don’t get too far off track to make things right before time’s up.”)

Box: Your Menu Depends on Ours QS does not run a private kitchen. Rest, care, help, quiet, and attention are shared resources. When others draw heavily from a shared channel, your felt menu can change too (in a trivial example: if everyone grabs the steak, you’re left with salad; the feeling analog is that some desired option loses its pull because the support for it is gone). The Law’s promise isn’t “you always get steak”; it’s “you will not be left without some feasible route to neutrality.” In practice that means substitutes appear, priorities shift with horizon, and ordinary coordination, scheduling, queuing, taking turns, becomes one of the Law’s principal tools. In short, our choices affect each other’s QS: we co-create each other’s menus in any shared environment, which is exactly what the next sections formalize.

5.1.8 Where we go next:

If QS is real, it should show up not as a purpose but as a narrowing of what is feasible from moment to moment. Next we make that concrete: 5.2 formalizes “choice sets” and “admissible policies,” showing how certain paths quietly drop out because they would make neutral closure infeasible.

5.2 Choice Sets and Admissible Policies

QS does not pick actions for you; it shapes the menu from which you pick. This section formalizes that menu, shows how it changes with the ledger and horizon, and translates the math into everyday felt experience and testable lab signatures. In social settings, menus can couple because many compensatory routes run through shared channels (for example, the same caregiver's time or the same pool of resources). Others' choices can tighten your menu, and your choices can tighten theirs. We therefore treat admissibility as conditionally social.

5.2.1 From thinkable to selectable

Let's define a hierarchy of option sets at time t for an agent:

- Thinkable set $\mathcal{U}(t)$: All actions, utterances, or omissions the agent could in principle imagine initiating at time t (for example, pressing send on a message, making a phone call, taking a nap, lashing out in anger, going for a walk, accepting help, revisiting a memory, seeking care). This is the unconstrained imagination set of what seems possible to do.
- Selectable set $\mathcal{S}(t) \subseteq \mathcal{U}(t)$: The subset of thinkable actions that come close to execution under standard decision processes, meaning those for which a premotor plan reaches threshold given current utility, conflict, and arousal conditions. In plain terms, these are options the agent is on the verge of doing under typical motivational and inhibitory controls. This roughly corresponds to what standard decision models would treat as available choices at that moment.
- Admissible set $\mathcal{A}(t) \subseteq \mathcal{S}(t)$: The subset of selectable actions that QS leaves realistically keepable, meaning pursuing them keeps the probability of neutral closure acceptably high given the running ledger $L(t)$ and horizon H_t . These are the options that make it through the QS filter. If an option would likely make neutral closure infeasible given the time left and current state, QS marks it inadmissible, so it will not arise with enough force to sustain, or it will fail to remain stable once initiated.
- Shared-feasible set $\mathcal{A}(t | R(t))$: The admissible set conditional on shared resource state $R(t)$. Here $R(t)$ represents external resources or contexts that many agents draw from (for example, available clinical slots, a partner's availability, quiet time, attention of others). When shared channels are saturated, $\mathcal{A}(t | R(t))$ shrinks or tilts toward lower-draw substitutes. For instance, if professional help is scarce at the moment, QS may favor a compensatory option that does not require that resource.

Intuition: Plenty is thinkable; less is selectable; only some is admissible. In group settings, even that admissibility can be jointly shaped by the environment.

5.2.2 A minimal decision-theoretic definition

Now we formalize the key construct Φ , including the shared-resource component.

Consider a candidate action $u \in \mathcal{S}(t)$, an option the agent is on the verge of taking under normal conditions. Pursuing u leads into a short-term policy fragment π_u , the sequence of steps and immediate consequences that follow from initiating u . Define a resource-aware feasibility-of-compensation score:

$$\Phi(u; L(t), H_t, R(t)) \equiv \Delta \Pr\{ L(T) \in [-K, K] | \pi_u, L(t), H_t, R(t) \} - \sum_r \lambda_r(t) \cdot \Delta r(u),$$

where the first term is the predicted change in the probability of neutral closure if u is pursued, and the second term is a shared-resource penalty.

In words, $\Phi(u; L, H, R)$ is the expected increase (or decrease) in the probability of a neutral final ledger if u is pursued, minus any costs u imposes on limited shared resources. Here:

- $R(t)$ is the vector of relevant shared resource levels or capacities at time t (for example, caregiver time, money, social bandwidth). Each resource r has capacity $C_r(t)$.
- $\Delta r(u)$ is the expected draw or usage of resource r by choosing u .
- $\lambda_{r,t}(t) \geq 0$ is the shadow price for resource r at time t , rising when a channel is crowded or near capacity, converting usage into a cost term $\lambda_r(t) \cdot \Delta r(u)$.

Using Φ , we define the conditional admissible set and weights:

$$\mathcal{A}(t | R) = \{ u \in \mathcal{S}(t) : \Pr\{ L(T) \in [-K, K] | \pi_u, L(t), H_t, R(t) \} \geq 1 - \varepsilon \}.$$

$$\omega(u; t) \propto \exp(\beta \cdot \Phi(u; L(t), H_t, R(t))),$$

with ε (tolerance for failure risk) set in advance, and $\beta > 0$ a gain parameter that increases as H_t shortens, making Φ differences more influential when time is scarce.

QS's action is to shrink the menu from $\mathcal{S}(t)$ down to $\mathcal{A}(t | R)$ and reweight options by $\omega(u; t)$ in favor of higher Φ . The agent's action is to choose or sample from $\mathcal{A}(t | R)$ according to the usual decision rule. QS shapes what is admissible; it does not dictate which admissible option you take.

5.2.3 How $\mathcal{A}(t | R)$ changes with ledger and horizon

Using the formalism above, we can describe three key effects in admissible-set dynamics:

- Ledger drift effect: If $L(t) \ll 0$, meaning recent net pain far outweighs pleasure, then repair-enabling or relief-providing options tend to have $\Phi > 0$ because they improve the chances of ending neutral. Those options move into $\mathcal{A}(t | R)$ and gain weight ω . Conversely, options that would deepen the debt, typically with $\Phi < 0$, slide out of \mathcal{A} , and you feel an aversion or “no go” for them.
- Horizon effect: As expected time H_t shrinks, β rises and the admissibility threshold tightens. This means $\mathcal{A}(t | R)$ becomes narrower and more reparative in content. With long horizons, QS is permissive because there is time to recover from detours. With short horizons, QS is conservative because little time remains, so it allows only highly compensable moves. The shorter your runway, the more strongly QS tilts you toward actions that keep neutrality feasible.
- Social coupling effect: When a shared channel saturates, the shadow price $\lambda_r(t)$ spikes. As a result, options with high $\Delta r(u)$ lose admissibility for most agents because their Φ is penalized heavily. An exception is agents with very short horizons, whose high β may preserve priority for critical uses despite the cost. When load eases, $\lambda_r(t)$ falls and those options become admissible again for more people.

These lead to clear empirical signatures: horizon scaling of behavior as H_t decreases, and menu co-movement under congestion as shared resources become scarce.

5.2.4 Examples (low-drama, realistic)

Let's ground this in everyday scenarios, each illustrating how QS might manifest without dramatic or supernatural overtones:

- The unsent message: You've had a frustrating day (ledger slightly negative) and it's late evening with a big meeting tomorrow (short horizon for recovery). You type a cutting text message intending to vent. Option: “Send the scathing text right now.” Here $\Phi < 0$ because sending it would likely lead to regret and relational damage that you might not have time to repair tomorrow. QS outcome: $u \notin \mathcal{A}(t | R)$. You feel a visceral hesitation; perhaps you delete the draft. The action does not feel keepable. Instead, a different option may surface, such as venting privately or going to bed.
- The surprising nap: You're slightly ahead on a work project (ledger a bit positive) but suddenly feel exhausted midday. Normally you'd power through, but today you strongly consider a nap. Option: “Take a 30-minute nap.” Under a long horizon, QS is permissive, so the nap is admissible but not heavily weighted. Under a short horizon, for example when now is the only chance to rest before back-to-back meetings, β is higher and the nap's restorative value can make Φ positive. QS

outcome: you feel a stronger pull to nap when the context makes rest critical for maintaining balance.

- However, if your horizon were effectively short (say you have back-to-back meetings later, meaning now is the only chance to rest), then β is higher and a nap's restorative value might make Φ positive. QS outcome: You feel an unusually strong pull to nap *if* the context makes rest critical for maintaining balance. If not, QS doesn't mind either way.
- The half-finished beer: It's the end of a rough week (ledger quite negative) and you pour a second beer. Normally you'd finish it, but halfway through you lose the taste. Option: "Finish this drink." If you're already feeling bad, another drink may push the ledger down, especially if it worsens sleep or mood, so Φ is likely negative. QS outcome: the beer suddenly feels unappealing. A warmer alternative surfaces, such as calling a friend or choosing a comfort show, something more likely to lift mood or at least not add damage.
- The overdue apology: You had a falling-out with a close friend months ago and pride blocked you. Now you've had a health scare (horizon subjectively shortened) and feel lonely (ledger drifting negative). Option: "Reach out and apologize." Here Φ is strongly positive because reconciliation could bring relief and increase the chances of neutral closure. QS outcome: what once felt impossible now feels obvious and urgent. The action moves into your admissible set and is heavily weighted.
- The stalled project: You're pursuing a risky multi-step project that might pay off or become a sunk cost. With deadlines far off (long horizon), you keep pushing. As time runs short and interim results look bleak, continuing becomes increasingly uncompensable use of time. Option: "Persist in this failing approach." As H_t shrinks, Φ for persisting becomes negative. QS outcome: an internal brake appears, and you shift efforts. What looks like procrastination may be admissibility tightening under short horizons.
- Notice that we often preempt such scenarios by cultivating routines or personal rules that keep our choices in safer territory. A rule like no screens or emails after 10 PM, or avoiding alcohol on weeknights, removes risky options from the menu ahead of time. In QS terms, good habits raise the odds that compensable actions enter $\mathcal{A}(t)$ in the first place. Chapter 21 explores how building habits and avoiding queue traps can be seen as training the admissible set to stay within LoF's guardrails.

5.2.5 Neural translation (preview of 5.3)

Before diving fully into neural details, here's a preview of how the above translates into brain signals:

- vmPFC/OFC (valuation): encodes something like standard Q-values, but under QS it includes a Φ term. Options with higher Φ should show a value boost in vmPFC beyond what immediate reward predicts.
- ACC and rIFG (control): implement braking and commitment thresholds. Under QS, we expect stronger inhibitory signals for actions with low Φ , especially as H_t decreases. Conversely, clearly reparative actions with high Φ may show lower conflict and lower braking, making them easier to initiate.
- Insula and autonomic circuits (interoception): track the felt weight of options via bodily signals. Under QS, interoceptive markers should track $\omega(u; t)$. Options in the admissible set feel right or necessary, whereas inadmissible ones feel off or effortful.
- Social penalty coding (ACC/rIFG): when an action draws on scarce shared resources, ACC and rIFG should register an additional signal proportional to $\sum_r \lambda_r(t) \Delta r(u)$. This means that even if you personally value an action, control systems may generate an extra cost signal when the same channel is crowded.

In sum, we test for a QS-residual proportional to $\Phi(u; L, H, R)$ after controlling for utility, conflict, arousal, and other standard factors.

5.2.6 Lab tasks to elicit admissible-set dynamics

We design specific experiments to provoke and measure QS effects:

- Horizon-manipulated go/withhold task: Participants make go/no-go decisions under different horizon framings, for example “plenty of opportunities later” versus “last chances.” We vary options’ Φ so some choices foreclose compensation. Prediction: rIFG and ACC show stronger inhibitory signals for low- Φ options when horizons are short.
- Repair versus indulgence choice task: Two options have similar immediate utility but differ in Φ . Prediction: vmPFC value signals favor the repair option beyond what utility alone predicts, and this effect scales with $|L|$ and with H_t^{-1} .
- Sequential persistence (stall) task: Participants engage in multi-step sequences simulating a long-shot policy. Some sequences have low Φ . We manipulate perceived horizon. Prediction: under short horizons, low- Φ sequences stall more often, and ACC cost signals ramp up leading to early termination.

- Shared-channel congestion paradigm: We create settings where a resource (help tokens or time with a helper) is limited and multiple participants choose actions that draw on it or not. We manipulate contention. Predictions: as contention rises, menus co-move away from high-draw options; participants with short horizons retain preferential selection of critical high-draw compensatory options; ACC and rIFG signals track $\sum_r \lambda_r(t) \Delta r(u)$ under crowding.

5.2.7 Naturalistic telemetry signatures

Outside the lab, if QS is real, its effects should surface in longitudinal or sensing data:

- Menu shrinkage near closure: As horizon shortens in real life, observable repertoires should narrow toward reparative, meaningful, or closure-oriented acts.
- Stickiness asymmetry: High- Φ actions should show greater persistence once initiated, whereas low- Φ actions should show higher abort rates, especially as horizons shorten.
- Social graph tilt: As H_t decreases, people should reallocate social energy toward supportive or reconciliation-oriented connections and away from adversarial relationships, measurable as shifts in communication patterns.
- Menu co-movement under scarcity: During shared crises or resource saturation, multiple individuals' menus should tilt in tandem toward alternatives, and broaden again when capacity returns. These tilts can be quantified via sensors, metadata, or EMA.

(All four are quantifiable given modern data; phone sensors, communication metadata, and EMA can track these with preregistered models.)

5.2.8 Multi-agent admissibility: shared-resource shadow prices

QS operates on individuals, but individuals co-create option spaces. When many agents pursue the same high-demand option, shared resources $R(t)$ deplete and institutional constraints can tighten. We can formalize coupling as follows:

Let person i have selectable set $S_i(t)$ and admissible set $\mathcal{A}_i(t)$. Then:

$$\mathcal{A}_i(t) = \mathcal{A}_i(L_i(t), H_i(t), R(t), P(t)),$$

where $R(t)$ captures resource capacity and $P(t)$ represents policy and institutional affordances. When $R(t)$ or $P(t)$ tighten, some actions drop out of many people's menus simultaneously. When resources expand or policies enable access, $\mathcal{A}_i(t)$ expands across groups.

QS still does not pick winners. It constrains admissible combinations of choices across agents so that, given everyone's ledgers and horizons, compensability remains feasible in aggregate. This can look dramatic, for example when coordinated policies expand care access, but in our framing it is affordance reconfiguration, not teleological orchestration.

Empirical predictions (multi-agent):

- Synchronized option withdrawal: During shared shocks, many agents report the same kinds of plans losing stickiness while relief options gain weight, after controlling for ordinary factors.
- Policy window effect: When policy increases compensability (for example, expanded access to palliative care), we predict measurable expansion of $\mathcal{A}_i(t)$ for vulnerable groups, including increased help-seeking and measurable affect improvements.
- Network gradient: Under scarcity, individuals most dependent on the scarce channel show the earliest and strongest menu pruning.
- Fail patterns: If shared shocks produce no detectable admissible-set shifts, no horizon-priority effect, or no rebound when resources free up, the social coupling model weakens.

5.2.9 Differentiating QS from common rivals

How do we distinguish QS effects from more mundane processes?

- Simple homeostatic rebound: adaptation predicts return toward baseline after extremes, largely independent of horizon. QS predicts that compensatory tilt depends on time remaining and compensability. If shifts strengthen specifically when H_t is short and recovery options exist, that favors QS over generic homeostasis.
- Pure utility maximization: a classical model predicts choices based on immediate expected utility or discounted reward only. QS predicts a Φ bias even when immediate utility is matched, and predicts earlier termination of low- Φ sequences under short horizons.
- Risk aversion: standard risk models penalize uncertainty broadly. QS selectively penalizes non-compensable risk. Under short horizons, QS may even promote risky actions if they are the only feasible route to repair.
- Other learning and fatigue processes: we include these as nuisance terms. QS is present if a $\Phi \times H$ interaction survives controls, along with social-penalty effects under congestion.

Empirical discriminator: The hallmark of QS is a $\Phi \times H$ interaction (and the λ_t social penalty effects) that survive controls. If after controlling for reward, risk, fatigue, adaptation etc., we still see that *shorter horizons amplify the influence of Φ on choices and signals*, and that including a social penalty term explains variance under congestion, then those are patterns not captured by simpler models. A rival would have to explicitly incorporate an equivalent mechanism to reproduce that.

5.2.10 Failure modes specific to $\mathcal{A}(t | R)$

Here are the concrete ways the QS mechanism on the menu level could fail. These align with the fail conditions listed earlier but stated in terms of the formalism:

- Admissible-set leakage: We observe, repeatedly, low- Φ trajectories proceeding unchecked near closure in well-measured cases where compensatory channels were in fact available. (E.g. someone in hospice with opportunities for reconciliation or comfort still goes on a destructive binge with zero internal resistance.) This would indicate QS failed to prune an obviously bad path.
- Null Φ -residuals: When we compute Φ for actions (using a conservative or low-bias feature approximation) and look at neural or behavioral data, we find no residual differences – i.e. nothing in brain signals or choice patterns correlates with Φ once standard factors are accounted for. If all ROIs show flat residuals, it means QS might be a ghost.
- No horizon interaction: Admissible menus do not systematically narrow or tilt as $H \downarrow$ (when compensatory actions exist). If having less time doesn't change behavior in our tasks (people act the same regardless of horizon length in terms of compensation vs. indulgence choices, etc.), then our core premise of “endgame tightening” is false.
- Rival fit is just as good: A constraint-free model (e.g. predictive coding + risk + fatigue, with no Φ or horizon terms) can fit the menu shrinkage, stickiness asymmetries, and value shifts *just as well*. If we can explain everything QS purports to explain without invoking a fairness constraint, then by Occam's razor we don't need QS.
- No social coupling: When we introduce measurable congestion (or observe it naturally) and include a $\sum_r \lambda_{rt} \Delta r$ regressor, we find no effect. Menus don't co-move when resources are scarce; ACC/rIFG show no special signal; behavior doesn't adjust collectively. That would mean admissibility is essentially independent, not conditional on shared context as we think.
- No horizon-priority under scarcity: In group settings where some have shorter horizons than others and resources are limited, if we do not see short-horizon

individuals retaining preferential access to high-draw options (they should be the last to drop those), then QS isn't implementing the expected fairness in order.

If these patterns are robustly observed and replicated, we would have to either retreat to seeing QS as a mere tendency or reject it entirely as a law-level mechanism.

5.2.11 Collective menus and scarcity coupling (practical note)

QS operates on individuals, but as highlighted, individuals' menus are entangled in shared conditions. In practical terms:

- For readers (intuitive take): When you notice that an option just “falls away” from your mind or that a reparative idea suddenly gains ease, that’s the inside feel of $\mathcal{A}(t | R)$ moving. And if you’re in a group crisis or high-demand situation, notice how your options seem to shift *along with everyone else’s*—that’s the social QS component you can practically feel in sync with your peers.
- For labs (implementation): To analyze this, one would compute person-specific Φ features using straightforward proxies (e.g. ReliefGain, RepairGain, HarmRisk, OptionFlexibility indices) and incorporate time-varying $\lambda_{rt}(t)$ covariates gleaned from objective data (e.g. queue lengths, wait times, pricing, indicators of contention). Then test for the QS signatures: the Φ residuals in brain and behavior, the $\Phi \times H^{-1}$ interactions, and co-movement of behaviors with shared λ_t fluctuations. Use appropriate models (GLMs, hierarchical Bayes) and robust cross-validation. Note: If analyzing count-like outcomes (e.g. number of options pursued), check for overdispersion (variance/mean > 1.2) and switch to a negative binomial model when needed to avoid misestimating effects.
- For clinicians (actionable advice): If QS is real, one can support it by widening shared channels that boost Φ . For example, ensure adequate pain control, sufficient sleep opportunities, access to reconciliation or counseling, and protected quiet times. In essence, make resources available that allow compensatory moves. QS cannot allocate what isn’t there – if the “steak” is gone, QS will force substitutions. By increasing supply (or policy support) for high- Φ actions (like making palliative care accessible), you are literally expanding patients’ admissible sets to include the very things that help them find balance.

Takeaway: QS converts “What can I do?” into “What can I sustain that keeps neutrality feasible?” In groups, the answer is a conditional admissible set $\mathcal{A}(t | R)$, a menu that tilts with your ledger, tightens with your horizon, and flexes with shared load.

5.2.12 Where we go next:

Mechanisms should touch everyday life. The most economical place to look is sleep. Section 5.4 treats dreams as potential low-cost counterweights that can change tomorrow's ledger without high real-world cost, linking overnight processing to the same feasibility logic that trims daytime options.

5.3 Neural Correlates: rIFG, ACC, vmPFC, Insula

The Queue System is a weighting layer, not a brand-new neural module. We expect its footprints to appear as modulations in already-known circuits for valuation, cognitive control, and interoception – specifically residual signals that correlate with our fairness constraint variables after accounting for ordinary decision influences. In this section, we specify the expected roles and signatures for four key regions/nodes: the right inferior frontal gyrus (rIFG), the anterior cingulate cortex (ACC), the ventromedial prefrontal cortex (vmPFC/OFC), and the insula/autonomic system. For each, we outline what it normally does, what QS predicts it should do (the QS-residual pattern), and how to model/test that. We include concrete model terms, example tasks, and fail conditions for each. Throughout, when we say “QS-residual,” we mean *the part of a signal uniquely explained by feasibility-of-compensation features (Φ and related terms) above and beyond standard predictors.*

Recent neuroscience evidence suggests the brain actively maintains an internal balance of value. For example, human fMRI studies show that the rostral anterior cingulate cortex (rACC) tracks imbalances among competing goals, while the ventromedial prefrontal cortex (vmPFC) signals the degree of corrective action a choice provides. In one experiment, participants had to keep two rewards in equilibrium; rACC activity rose with any growing disparity, and vmPFC activity reflected the amount of “redress” each decision offered (Juechems, 2019). This implies a neural network dedicated to value equilibrium, directly supporting the Law of Fairness: the brain appears wired to detect and counteract deviations to keep the experiential ledger balanced.

5.3.1 vmPFC/OFC — Value with a Compensation Term

Role: vmPFC/OFC integrates diverse evidence (reward, context, internal states) into a subjective value signal that guides choice. It is thought to encode an integrated subjective value signal – essentially a ‘common currency’ – for comparing different options. **QS prediction:** Under QS, vmPFC should encode a value boost for options that increase the probability of neutral closure – that is, options with high $\Phi(u; L, H, R)$ will register as more valuable than their immediate utility alone would suggest. Conversely, options that threaten future compensation (low or negative Φ) will see value tempered in vmPFC.

Model: We include a Φ term in the GLM for vmPFC’s BOLD signal. For example:
 $BOLD_{vmPFC}(t) \leftarrow Utility(u) + Conflict/Cost(u) + Arousal(t) + \gamma_1 \cdot \Phi(u; L, H, R) + \varepsilon.$

Here γ_1 is the coefficient on the QS term.

Prediction: $\gamma_1 > 0$. In other words, the Φ term should have a positive weight in vmPFC: higher Φ (more compensatory options) leads to higher neural value signals. Additionally, this Φ influence likely scales with $|L|$ (ledger imbalance magnitude) and with H^{-1} (greater emphasis when the horizon is short). We might see an interaction where vmPFC is especially responsive to Φ when the need is great or time is short.

Task signatures:

- In a repair vs. indulgence choice with matched immediate utility, we expect vmPFC activity (or chosen value) to favor the *repair* option when $\Phi_{\text{repair}} > \Phi_{\text{indulgence}}$. Moreover, this bias should strengthen as the horizon shortens or as the individual's ledger becomes more negative.
- In an ambiguous-value gambles task where two gambles have the same EV but differ in "option flexibility" or reversibility (one gamble might allow you to bail out halfway with minimal loss – thus more compensable if it starts going bad), we'd predict vmPFC prefers the more flexible gamble when ledger drift $|L|$ is large. Essentially, vmPFC should implicitly value the *ability to compensate later*, not just the raw EV, reflecting QS influence.

Null/fail scenario: If, after nuisance modeling, $\gamma_1 \approx 0$ (vmPFC) shows no extra sensitivity to Φ , or if any apparent Φ -effect vanishes once we control for known factors like risk, ambiguity, or habit, then QS isn't contributing to vmPFC value coding as expected. For example, if what we thought was a Φ -effect was really just risk aversion, adding a risk regressor would erase it – that would be a fail for QS's unique role.

Causal evidence is consistent with a prefrontal enforcement role: low-frequency TMS disrupting right DLPFC reduces people's willingness to reject unfair offers in the Ultimatum Game while leaving perceived unfairness judgments intact (Knoch 2006). This finding supports the involvement of executive control circuits in norm-consistent action selection. While it does not directly test horizon-sensitive compensability encoding, it aligns with the broader claim that prefrontal control systems can implement constraint-based behavioral regulation of the kind QS would require.

5.3.2 ACC — Cost Monitoring, Policy Viability, and Social Penalty

Role: The ACC (particularly dorsal ACC) monitors expected control costs, conflict, errors, and (we posit) the viability of ongoing policies. It often signals when things are difficult, unexpected, or when a considered path might be problematic.

QS prediction: Under QS, ACC should track two specific things: (i) the shadow price for non-compensability as the horizon shortens (basically an urgency or "stakes" signal

when time is short and an action has big future impact), and (ii) shared-resource penalties during congestion (when an action's pursuit is imposing a cost on others or likely to fail due to external crowding).

The same anterior insula-rostral ACC circuit is engaged by personal and vicarious pain. Singer (2004) showed bilateral anterior insula and rostral ACC activate when participants experience pain and when they observe a loved one in pain. This “shared affect” pathway provides a neural basis for automatic sensitivity to others’ costs, matching the QS prediction that ACC should register social-penalty terms under congestion.

Model: We augment an ACC GLM with a horizon-interaction term and a social penalty term:

$$\text{BOLD_ACC}(t) \leftarrow \text{Conflict}(u) + \text{Error/PE}(t) + \text{Arousal}(t) + \gamma_2[\Phi(u; L, H, R) \times H^{-1}] + \gamma_3 \sum_r \lambda_{rt}(t) \Delta r(u) + \varepsilon$$

Here γ_2 captures how ACC responds to Φ when combined with a short horizon ($\Phi \times H^{-1}$), and γ_3 captures sensitivity to the social resource penalty (the summed $\lambda \Delta r$).

Predictions: $\gamma_2 > 0$ and $\gamma_3 > 0$. That is, ACC activity increases for actions with a given Φ as the horizon shrinks (horizon scaling of the compensation-urgency signal), and ACC increases with the social penalty term (when an option costs a lot of shared resource, ACC registers that). In effect, ACC should be more activated by a risky or uncompensable move if time is almost up (because it “knows” this could be catastrophic), and also by attempting something when others need the same resource (signaling social cost or conflict).

Task signatures:

- In the sequential persistence task mentioned earlier: as a participant goes down a multi-step path that looks less and less likely to compensate the ledger, ACC should show a ramping signal indicating increasing cost or conflict, and this ramp should be *steeper under short horizons*. In behavior, this could correspond to the person feeling compelled to stop sooner when time is short (stalling out, guided by ACC’s growing signal).
- In a congestion paradigm with limited shared help tokens, even if a participant’s subjective utility for an action is held constant, ACC should scale with $\sum_r \lambda_r \Delta r$. For example, two tasks might be equally rewarding to me, but one uses up a lot of a shared resource that’s scarce (like teacher time in a classroom). ACC would fire more for considering that task under scarcity, reflecting an internalization of the

external constraint (“it *feels* harder or more fraught to choose that because it’s socially costly”).

Null/fail scenario: If ACC’s effects reduce to classic conflict or error signals only – for example, if γ_2 and γ_3 coefficients are indistinguishable from zero after proper controls – then QS isn’t adding explanatory power. We’d conclude ACC is not showing any special horizon or social sensitivity beyond well-known functions. A concrete fail would be: manipulate time pressure and congestion, but ACC BOLD and decisions remain fully explained by basic difficulty or subjective effort, with no extra term needed.

5.3.3 rIFG / Basal Ganglia — Admissibility Brakes and Commit Gates

Role: The right inferior frontal gyrus (rIFG) is a key node for inhibitory control (“stop” signals), often in concert with the subthalamic nucleus (STN) and globus pallidus interna (GPi) in the basal ganglia which implement the braking and gating of actions. These circuits decide whether to go, hold, or stop an action.

QS prediction: rIFG/STN should show raised brakes for low- Φ actions (especially those that would worsen the ledger without time to fix) and lowered brakes (or quicker gates) for high- Φ reparative actions, with these effects *more pronounced when horizons are short*. In short, QS biases the inhibitory threshold depending on compensability.

Model (trial-level):

$$\text{BOLD_rIFG/STN}(t) = \text{StopDemand}(u) + \text{Conflict}(u) + \alpha_4 \Phi(u; L, H, R) + \alpha_5 [\Phi(u; L, H, R) \times H^{-1}] + \varepsilon, \text{ where } \alpha_4 < 0 \text{ and } \alpha_5 < 0 \text{ (so higher } \Phi \text{ leads to lower activation, i.e. more braking on low-} \Phi \text{ actions, especially as time shortens)}$$

We include standard predictors like any stop-signal requirement or conflict, then a term for $-\Phi$ (so that a positive γ_4 means greater activity for lower Φ) and an interaction $-\Phi \times H^{-1}$. Predictions: $\alpha_4 < 0$ and $\alpha_5 < 0$ (meaning rIFG/STN activity will increase as Φ decreases, especially when the horizon is short). Concretely, inhibitory regions fire more strongly when Φ is low (that is, γ_4 effectively > 0 for low Φ) and even more so when time is short (γ_5 adds extra braking for low- Φ under short H). This means rIFG/STN are extra suppressive of actions that aren’t compensable, especially near endgame.

Behaviorally, we’d also expect to see coupling in parameters of decision models: e.g. in a drift-diffusion model of decision-making, *boundary height* might increase for low- Φ options (harder to commit to them), or starting point might bias toward not doing low- Φ actions when time is short, etc.

Behavioral coupling: Yes, using sequential sampling models we can look at whether response thresholds get modulated by Φ . If QS is active, in conditions where an option

has a low Φ , people might require more evidence or take longer to act on it (like an internal hesitation). As H shortens, this hesitation becomes more pronounced (for low- Φ) or perhaps flips to a quick commitment for high- Φ (like a starting point shift toward the high- Φ alternative). These would appear as DDM fits showing threshold or bias changes correlated with Φ and H .

Null/fail scenario: If we find no modulation of stopping or going behavior by Φ after accounting for conflict and risk, then QS isn't showing up in inhibitory control. For instance, if participants' stop-signal reaction times or go/nogo accuracy do not differ between compensable vs. non-compensable contexts (controlling for difficulty), or rIFG activation is entirely explained by generic difficulty, that's a failure for QS. No $\Phi \times H$ effect on response thresholds would likewise indicate QS is not influencing motor gating.

5.3.4 Insula / Brainstem / Autonomics — Somatic Markers of the Menu

Role: The insula (particularly anterior insula) integrates interoceptive states (heart rate, gut sensations, arousal) and contributes to the subjective feeling of bodily and emotional states. It often signals when something feels “off” versus “okay” and is central to somatic marker signaling.

QS prediction: The insula and associated autonomic outputs should encode a bodily tilt reflecting the admissible set. High- Φ actions should feel easing or appropriate, whereas low- Φ actions should feel effortful, tense, or dissonant, especially under conditions of short time or high stakes. In this framing, gut-level signals align with QS weighting: when an option is inadmissible, it fails to sit right; when it is admissible and compensatory, it carries a felt sense of resonance or relief. Critically, this differentiation must persist after controlling for generic arousal and valence.

Model: Because insula integrates bodily signals, we incorporate the computed weight $\omega(u; t)$ directly:

$$\text{BOLD_insula}(t) = \text{Arousal}(t) + \text{Valence}(t) + \gamma_6 \omega(u; t) + \varepsilon,$$

with $\omega(u; t) \propto \exp(\beta \Phi(u; L, H, R))$.

Here $\omega(u; t)$ reflects the QS-weighted salience of option u . In practice, ω is derived from the fitted decision model (e.g., softmax weights), and we test whether insula activity correlates with ω after accounting for arousal and valence.

Physiological coupling: Peripheral measures provide an additional test. High-frequency HRV (vagal tone), skin conductance level (SCL), respiratory sinus arrhythmia, and related autonomic indices should covary with ω . When an agent considers or chooses a high- ω (compensable) option, we predict relative physiological ease (e.g., increased vagal tone,

lower sympathetic arousal) compared to low- Φ options, even when overall excitement or unpleasantness is matched. The claim is not that the body “knows fairness,” but that interoceptive systems track QS-weighted feasibility rather than raw valence alone.

Null/fail scenario: If interoceptive and autonomic changes track only gross emotional valence or stress and show no relationship to w or Φ after proper controls, then QS lacks a somatic marker component. A concrete failure would be indistinguishable heart rate, SCL, and insula responses for two equally arousing options when one is strongly non-compensable and the other strongly compensable.

5.3.5 Joint Signature: The QS-Residual After Nuisance Controls

It's important to emphasize the joint nature of the QS prediction: it's not just one region or measure, but a pattern across multiple signals. To claim evidence of QS, we look for a convergent residual signature across modalities *after controlling for all known nuisances*.

Concretely, we fit multivariate models that include the standard predictors (utility, conflict, risk, ambiguity, habit, fatigue, arousal, etc. – see 5.5 for exhaustive list) and then test whether adding the Φ -based terms yields reliable improvements.

One way to conceptualize it is the “stack” of residuals per trial or event:

$$\text{Signal}(t) = \text{Utility} + \text{Conflict} + \text{Arousal} + \text{Habit} + \text{Risk} [\text{nuisance backbone}] + \gamma_{\Phi} \Phi(u; L, H, R) + \gamma_H [\Phi \times H^{-1}] + \gamma_S \sum_r \lambda_{rt} \Delta r(u)$$

We do this for various signals (neural ROI activity, behavioral outcomes like choice or RT, etc.). Then we report: (i) the region-wise γ estimates with uncertainty intervals, (ii) the out-of-sample predictive improvement from including those terms (e.g., difference in LOO-IC or WAIC, reduction in Brier score or log-loss for predicting choices), and (iii) robustness checks (does it hold if we tweak the nuisance set or use alternative measures?).

The critical question: Do the QS terms collectively explain residual variance that the nuisance model cannot, and do they generalize to new data?

Null/fail scenario: If adding the QS terms does not improve prediction (model fit) and the coefficients for these terms straddle zero with tight confidence intervals across datasets, then we have no evidence for QS. For example, if our cross-validated log-likelihood or LOO-IC is basically unchanged by including Φ , H^{-1} , and social penalty terms, and their posteriors all overlap zero, that's a pretty clear *failure to detect QS*.

(This joint test is crucial – any single result can be a fluke, but a pattern across brain, behavior, sleep, etc., with consistent Φ effects, is harder to dismiss.)

5.3.6 Horizon Scaling and Endgame Compression

A core prediction of the Law of Fairness is what we call horizon scaling: as the end of the line approaches ($H_t \rightarrow 0$), QS effects should intensify. In neural terms:

- vmPFC: should show a larger Φ -related boost for compensatory (repair/relief) choices when horizons are short. With long horizons, there should be smaller or no additional Φ -related boost; with short horizons, a stronger Φ sensitivity.
- rIFG/STN: should show a steeper inhibitory response to low- Φ options as the horizon shortens. With ample time, the brain may tolerate a risky or low-compensability idea; with little time remaining, inhibitory control should increase for that same option.
- ACC: should display greater cost signals for uncompensable continuations when time is short, and larger social-penalty coding during congestion under short horizons. As H_t decreases, the cost of non-compensable or socially costly paths should scale upward in ACC activity.
- Insula/autonomics: should show sharper somatic differentiation between high- ω (admissible, compensatory) and low- ω (inadmissible, non-compensatory) options when time is short. In effect, interoceptive contrast between viable and non-viable paths should increase as $H_t \rightarrow 0$.

At a population level, across individuals, we also anticipate endgame compression as an empirical signature of drift regularization. As different people approach death of mind ($H_t \rightarrow 0$), the cross-sectional variability in their cumulative ledgers $L(t)$ should decrease, conditional on intact compensatory channels and adequate measurement. This compression is not definitional; it is a predicted dynamical consequence of horizon-sensitive drift.

More precisely: we predict a measurable reduction in cross-sectional ledger variance as H_t decreases, provided that compensatory routes remain available. In contexts where compensation is feasible, late-stage cumulative ledgers should cluster more tightly around neutrality than mid-life ledgers.

Null/fail scenario: No systematic $\Phi \times H_t$ interactions anywhere they should be. For example, experimentally manipulating deadlines yields no change in QS-related neural or behavioral signals; or hospice versus earlier-stage patients show no difference in convergence patterns. If near-death ledgers remain as dispersed as mid-life ledgers under good measurement and intact channels, there is no evidence of compression.

Operationally, we require at least modest compression under adequate data quality, defined as: final ledger mean within ± 0.15 SD (standardized HCU units), terminal drift statistically indistinguishable from zero over a preregistered window, and cross-sectional variance ratio ≤ 0.8 relative to mid-life stages. If these conditions are not met under well-measured circumstances where compensation is possible, the Law's horizon-scaling prediction is not supported.

This section therefore treats compression not as an assumption but as a measurable, falsifiable dynamical signature of the QS mechanism.

5.3.7 Social Coupling: Neural Correlates of Shared-Resource Shadow Prices

When people share a limited resource, QS predicts a socially coupled admissibility structure. In formal terms, the shadow prices $\lambda_r(t)$ associated with shared resources $R(t)$ enter $\Phi(u; L, H, R)$ through the penalty term $\sum_r \lambda_r(t) \Delta r(u)$, and therefore alter $\mathcal{A}(t | R)$. Under measurable congestion, these shadow prices should increase and shift neural and behavioral signals in structured, not purely stress-driven, ways.

- ACC/rIFG: increased control cost signals and stronger inhibitory responses for high-draw options. If an action consumes a scarce resource (large Δr), ACC and rIFG activity should scale with the penalty term $\sum_r \lambda_r(t) \Delta r(u)$, over and above generic conflict or task difficulty.
- vmPFC: partial erosion of value for high-draw options under scarcity, unless horizon priority applies. For example, if a hospital bed is scarce and H_t is long, vmPFC value coding for occupying that bed should decrease relative to baseline; if H_t is short (horizon priority), this devaluation should attenuate. This produces a measurable $\Phi \times H_t$ interaction within valuation signals.
- Insula/autonomics: greater interoceptive friction when considering high-draw actions during congestion. This should manifest as increased anterior insula activity and sympathetic markers (e.g., elevated SCL) proportional to $\sum_r \lambda_r(t) \Delta r(u)$, independent of global stress covariates.

Empirically, we test whether measured congestion (queue length, wait time, capacity utilization, policy constraints) predicts structured shifts in $\mathcal{A}(t | R)$ and corresponding neural residuals. Specifically, we examine whether:

- Menus co-move across individuals as $\lambda_r(t)$ rises.
- ACC/rIFG signals track the social penalty term $\sum_r \lambda_r(t) \Delta r(u)$.
- vmPFC value coding reflects horizon-prioritized devaluation under scarcity.

Null/fail scenario: If neural and behavioral signals show no systematic relationship to measured congestion after controlling for difficulty, effort, and global stress, then the social-coupling component of QS is unsupported. In particular:

- If including $\sum_r \lambda_r(t) \Delta r(u)$ does not improve predictive performance (no change in LOO-IC/WAIC and coefficients straddle zero),
- If menus do not co-move under scarcity,
- If all observed effects reduce to undifferentiated stress responses,

Then admissibility appears independent rather than socially conditioned, and the shared-resource extension of QS fails.

Critically, the prediction is not “stress changes behavior,” but “structured shadow prices alter admissible sets in proportion to measurable congestion.” The distinction is testable via explicit $\lambda_r(t)$ modeling and cross-validated residual analysis.

5.3.8 Measurement Plan: Modalities, Contrasts, and Power

To detect subtle QS effects, we require a multimodal measurement strategy with preregistered contrasts, cross-validated models, and explicit nuisance controls aligned with Sections 5.2–5.3. The goal is not single-region activation but convergent Φ - and horizon-sensitive residuals across neural, behavioral, and physiological channels.

- fMRI: Use event-related designs that isolate decision onset, commitment, and stopping phases. Predefine regions of interest (ROIs) in vmPFC/OFC, dorsal and rostral ACC, rIFG, STN (if resolvable), and anterior/mid-insula based on anatomical masks or subject-specific localization where feasible. Apply family-wise error correction within the ROI set and control false discovery across contrasts. Preregister regressors including $\Phi(u; L, H, R)$, H^{-1} , $\Phi \times H^{-1}$, and $\sum_r \lambda_r(t) \Delta r(u)$, alongside the full nuisance backbone (utility, conflict, risk, ambiguity, habit strength, fatigue, arousal). Model comparisons must use cross-validated metrics (e.g., LOO-IC or WAIC), not in-sample fit alone.
- EEG/MEG: Examine time–frequency signatures associated with valuation and control. Frontal midline theta (ACC-related conflict monitoring), beta-band modulation (motor inhibition, rIFG/STN coupling), and late positive components linked to valuation should be tested for Φ and $\Phi \times H$ interactions using single-trial regression. Temporal specificity is critical: QS effects should appear during evaluation and gating windows, not uniformly across the trial.
- Intracranial/ECoG (when available): In clinical settings with depth electrodes, assess high-gamma activity in vmPFC, ACC, rIFG, and insula during horizon-

manipulated and compensability tasks. High-gamma provides a closer proxy to local population firing and can validate hemodynamic findings.

- Peripheral physiology: Record high-frequency HRV (parasympathetic tone), skin conductance level and responses (sympathetic arousal), and respiratory sinus arrhythmia. Test whether $\omega(u; t)$ or $\Phi(u; L, H, R)$ predicts physiological differentiation after controlling for valence and global arousal. Effects must survive inclusion of stress covariates.
- Behavioral modeling: Fit sequential sampling models (e.g., drift–diffusion models) with Φ and $\Phi \times H^{-1}$ terms entering boundary height, starting point bias, or drift rate, depending on theoretical mapping. Confirm that including these terms improves out-of-sample prediction of choice and response time.
- Power considerations: Conduct simulation-based power analyses (Section 5.5.9) using synthetic agents with known Φ and horizon parameters. Anticipated standardized effect sizes are modest ($\beta \approx 0.10\text{--}0.20$). Within-subject designs will likely require approximately 60–100 sessions for stable detection of $\Phi \times H^{-1}$ interactions; between-subject comparisons may require $N \geq 120$ depending on variance structure. Target statistical power ≥ 0.90 for primary interaction terms, with preregistered stopping rules and alpha control where applicable.

Converging evidence across modalities is essential. Detection of QS requires (i) significant Φ -based residual terms after nuisance controls, (ii) cross-validated improvement in predictive metrics, and (iii) robustness across measurement channels. Isolated findings without replication across modalities will not count as confirmation.

Failure criterion: If preregistered Φ , $\Phi \times H^{-1}$, and $\sum_r \lambda_r(t)\Delta r(u)$ terms fail to improve predictive performance across modalities, and coefficients consistently straddle zero with narrow confidence intervals under adequate power, the measurement plan yields no evidence for QS.

Careful design—including blocked cross-validation, counterbalancing of horizon manipulations, and strong nuisance controls—ensures that any detected QS effect reflects structured compensability dynamics rather than stress, difficulty, or generic decision noise.

5.3.9 Confounds and How We Guard Against Them

We recognize that QS-like patterns could be mimicked by well-known cognitive and affective processes. The following controls are therefore mandatory. A QS effect is credible only if it survives this confound audit.

- Reward confounding: High- Φ options might coincide with higher immediate reward. In key contrasts (e.g., repair vs. indulgence), we match immediate utilities and include trial-wise utility regressors in all models. If Φ effects disappear once immediate reward is controlled, the result is attributed to standard value coding, not QS.
- Risk and uncertainty: Compensable options may correlate with lower variance or entropy. We include risk metrics (variance, entropy, ambiguity indices) as regressors and construct tasks where options are matched on risk but differ in compensability. Persistence of Φ effects after risk controls is required to claim QS involvement.
- Global arousal and stress: Short horizons and congestion can elevate general stress. We record skin conductance, heart rate variability, and related autonomic markers and include global arousal regressors in neural and behavioral models. A genuine QS term must improve model fit beyond a pure arousal explanation and must show structured $\Phi \times H$ interactions rather than uniform stress amplification.
- Fatigue and time-on-task: Horizon manipulations can correlate with later trial positions. We counterbalance order, randomize block sequences, and include trial number, block number, and habit-strength covariates. Any apparent “stalling” or pruning must remain after these controls.
- Demand characteristics: Participants may infer expected behavior under “short time” instructions. Horizon manipulations are embedded in cover contexts (e.g., financial vs. temporal framing), and belief checks are administered post-task. Participants who do not endorse the horizon manipulation should not show $\Phi \times H$ effects; this serves as a manipulation-validity check.
- Baseline adaptation: Generic opponent-process or homeostatic rebound can produce balancing patterns independent of QS. We include baseline mood drift and autoregressive terms in longitudinal models (see Section 3.6). Φ -based terms must add predictive value beyond these standard adaptation dynamics.
- Social stress vs. structured congestion: In shared-resource paradigms, increased stress from crowding could mimic QS social-penalty effects. We therefore include independent stress indices and test whether $\sum_r \lambda_r(t) \Delta r(u)$ terms predict behavior or neural signals after stress is controlled. If congestion effects reduce to generic stress, the QS social-coupling claim fails.
- Model overfitting: Because Φ and $\Phi \times H$ terms are theory-driven, we require out-of-sample validation (cross-validated log-likelihood, LOO-IC, or WAIC improvement). In-sample significance alone does not count as evidence.

In summary: if QS-related terms (Φ , $\Phi \times H^{-1}$, and $\sum_r \lambda_r(t) \Delta r(u)$) do not improve prediction beyond nuisance regressors, or if their coefficients consistently collapse toward zero under rigorous controls, then the QS account weakens accordingly. We treat any apparent QS detection as provisional until it survives this confound battery.

5.3.10 Clinical and Everyday Relevance

Though this chapter is technical, the stakes are practical. If QS operates as described, its effects should appear not only in laboratory tasks but in clinical settings and ordinary life.

- Palliative settings: QS predicts that, as H_t shortens, vmPFC valuation should tilt toward options that increase compensability (e.g., reconciliation, meaningful conversation, adequate pain relief), while rIFG/ACC should increasingly inhibit futile or non-compensable medical escalations. Insula signals may reflect a felt shift toward what is “right-sized” given the horizon (for example, declining another round of burdensome intervention that offers little compensatory benefit). These are testable claims: we would expect measurable $\Phi \times H$ interactions in decision behavior and, where feasible, neural or physiological proxies. If no horizon-dependent tilt appears in well-measured end-of-life cohorts with intact channels for compensation, the QS account weakens.
- Addictions and compulsions: These contexts involve strong internal drives that can temporarily dominate valuation. QS predicts that when state and horizon cues are salient (e.g., health scare, legal consequence, family rupture), inhibitory control and valuation circuits should show increased sensitivity to compensability. Concretely, rIFG/ACC activity and choice thresholds should shift in the direction of braking low- Φ actions, particularly under shortened perceived horizons. This does not imply guaranteed recovery; rather, it predicts measurable increases in compensability-sensitive control signals when credible horizon compression occurs. If such horizon-contingent shifts are absent under well-controlled conditions, QS lacks support in this domain.
- Workplaces and educational settings: Under high congestion (shared-resource strain, time scarcity), QS predicts structured menu co-movement. Individuals should simultaneously deprioritize low- Φ activities and reweight toward actions that preserve flexibility or prevent downstream cost. In neural or behavioral telemetry, this would appear as synchronized changes in option weights correlated with measurable congestion indices (e.g., workload density, scheduling bottlenecks). Importantly, if behavior shifts are fully explained by

generic stress without a structured $\Phi \times H$ or shared-penalty component, then the effect is not uniquely attributable to QS.

In all cases, the prediction is conditional and measurable: QS effects should scale with ledger state $L(t)$, horizon H_t , and shared constraints $R(t)$. Where channels for compensation are blocked or measurement quality is poor, we expect ambiguity rather than forced convergence.

In practical terms, making the implicit explicit—recognizing how option menus narrow or tilt under ledger imbalance and time pressure—can inform compassionate intervention. Expanding compensatory channels (rest, reconciliation, access to care, protected time) should enlarge the admissible set $\mathcal{A}(t | R)$. If QS is real, widening feasible routes to balance should produce observable shifts in valuation, control signals, and behavior. If no such shifts occur despite improved channels and credible horizon compression, the mechanism requires revision.

5.3.11 What Would Falsify the Neural Story

To conclude this section, we state clearly what findings would falsify the QS-based neural account (and thus weaken the mechanism proposed for the Law).

- No Φ -residuals after rigorous controls. Across tasks, regions, and modalities, if adding $\Phi(u; L, H, R)$, $\Phi \times H^{-1}$, and shared-penalty terms does not improve model fit beyond a fully specified nuisance backbone (utility, conflict, risk, ambiguity, habit, fatigue, arousal, etc.), and the associated coefficients are tightly centered on zero across datasets, then QS has no detectable neural or behavioral footprint.
- No horizon interaction anywhere it should appear. If shortening H does not systematically amplify Φ -related effects in vmPFC (valuation), rIFG/STN (inhibitory control), ACC (cost/viability), or insula/autonomic signals, then the core horizon-scaling prediction fails. In particular, if well-powered manipulations of time-to-closure produce flat $\Phi \times H$ interactions, the “endgame tightening” mechanism is unsupported.
- No social-penalty coding under measurable congestion. If, in controlled scarcity paradigms or natural congestion contexts, ACC/rIFG activity and behavior do not covary with the shared-resource penalty term $\sum_r \lambda_{rt} \Delta r(u)$ after controlling for stress and generic task difficulty, then the social-coupling component of QS is not supported.
- Rival models match all signatures without Φ . If a constraint-free model (e.g., predictive coding or reinforcement learning with risk, ambiguity, fatigue, and stress terms) reproduces the full pattern of observed effects—including apparent

horizon scaling and any endgame compression—without requiring Φ or shared-resource shadow prices, then QS is not a necessary explanation. In that case, parsimony favors the simpler account.

- No endgame compression under adequate measurement. In longitudinal or end-of-life cohorts with intact compensatory channels and high data quality, if cumulative ledger variance does not decrease as $H \rightarrow 0$ and final ledgers do not cluster within the preregistered $\pm K$ band more tightly than mid-life distributions, then the compression prediction fails.

Any of these outcomes, if robust and replicated, would force us to revise or abandon the QS mechanism. The standard is not “some hint of effect,” but consistent, generalizable evidence across modalities and contexts.

If the predicted signatures repeatedly fail to appear under well-powered designs and strong controls, the responsible conclusion is not that the mechanism is hidden, but that it is absent or not law-level.

5.3.12 Where we go next:

Mechanisms should connect to everyday processes. The most economical next test bed is sleep. Section 5.4 examines dreams as potential low-cost counterweights—overnight processes that could adjust tomorrow’s ledger without heavy real-world expenditure—linking off-line processing to the same feasibility logic that trims daytime options.

5.4 Dreams as Low-Cost Counterweights

If the Queue System (QS) is a constraint enforced by ordinary neural machinery, sleep is its ideal workshop. Dreaming offers an offline, low-cost environment to nudge one's affective state back toward neutral without expending scarce waking resources such as time, money, social capital, or clinical interventions. This section explains how dreams may function as low-cost counterweights, how targets are selected, what neuromodulatory mechanisms are consistent with this account, what signatures we should measure, and what findings would falsify the claim.

5.4.1 Why Dreams Are Uniquely Suited to Compensation

- Minimal draw on shared resources: During sleep the body is largely immobile, socially offline, and metabolically efficient relative to waking. Counterweighting here does not consume social penalties $\sum_r \lambda_r \Delta r$ in the way waking actions might. In short, dreams can simulate repair or relief without incurring shared-resource costs.
- Plasticity window: REM and certain NREM stages are associated with memory replay, emotional reconsolidation, and extinction learning. This makes sleep a plausible substrate for reweighting affective associations. Experiences can be reprocessed under altered neuromodulatory conditions, potentially shifting their contribution to the running ledger $L(t)$ without requiring overt behavioral intervention.
- High internal controllability: In dreams, QS can sample, recombine, and manipulate memory traces and imagined scenarios without negotiating external constraints. The sleeping brain can freely recombine experiences. Upon waking, these reweighted traces may manifest as reduced fear responses, softened affect, renewed motivation, or altered salience of options, consistent with compensatory work having occurred offline.

5.4.2 The Counterweight Hypothesis (Formal Sketch)

Let ΔL_{day} denote the net ledger change during a given day (net relief minus net pain). Let D_{night} denote dream affect that night, operationalized as an integrated latent affect estimate derived from self-reported dream valence and physiological proxies (e.g., REM density, autonomic indices).

The core prediction is that dream affect counterbalances recent ledger drift in a horizon-sensitive manner:

$$E[D_{\text{night}} | \Delta L_{\text{day}}, H, R] = -a(H) \cdot \Delta L_{\text{day}} + \eta,$$

where:

- H is the current time horizon (remaining life expectancy or context-specific closure horizon).
- R represents relevant shared-resource conditions.
- $\alpha(H) \geq 0$ is a scaling function that increases as H shortens (i.e., $\alpha'(H) < 0$ when H is measured as time remaining).
- η is mean-zero noise.

Interpretation: After a negative day ($\Delta L_{\text{day}} < 0$), expected dream affect shifts positive in proportion to the magnitude of that deficit, and more strongly when the horizon is short. After an unusually positive day, dream affect may tilt negative (or dampen positivity) to prevent sustained overshoot.

Importantly, this is not a claim that every night perfectly counterbalances the day. The claim is statistical and conditional: over many nights and individuals, dream affect should show an inverse coupling to prior-day ledger drift beyond what baseline adaptation, circadian effects, or generic stress responses predict.

Empirical signatures

- Inverse coupling: Across nights, dream affect should correlate negatively with prior-day ΔL_{day} after controlling for baseline mood, trait affect, sleep quality, and circadian phase.
- Horizon modulation: The magnitude of this inverse coupling ($|\alpha(H)|$) should increase as H decreases. For example, individuals in short-horizon contexts (e.g., advanced illness or salient deadlines) should show stronger dream counterweight effects than long-horizon controls.
- Specificity beyond arousal: The effect should persist after controlling for daytime stress levels and sleep fragmentation, indicating it is not merely stress rebound.

Null/fail scenarios:

- No inverse coupling: If dream affect shows no systematic relationship to prior-day ledger drift after adequate controls and sufficient power, the counterweight hypothesis fails.
- No horizon interaction: If $\alpha(H)$ does not scale with H in contexts where compensatory channels remain intact, then the endgame modulation component fails.

- Full explanation by baseline models: If adaptation, generic affective rebound, or circadian rhythm fully account for observed dream patterns without requiring ΔL_{day} -dependent terms, then QS adds no explanatory value.

These conditions are necessary to keep the dream counterweight hypothesis within a falsifiable, mechanistic framework.

5.4.3 Mechanism: Neuromodulators and Circuitry

Several established features of sleep neurobiology plausibly support counterweight functions by creating conditions favorable for emotional reprocessing rather than immediate action:

- REM sleep (high ACh, low NE/5-HT): During REM, acetylcholine levels are elevated while norepinephrine and serotonin are markedly reduced. This neurochemical profile promotes plasticity while dampening noradrenergic “alarm” signaling, allowing emotionally salient material to be reactivated without full fight-or-flight engagement. Nightmares and intense emotional dreams can therefore occur without triggering sustained sympathetic escalation. Under the QS hypothesis, REM provides a window in which threat memories or unresolved affect can be represented in modified form, updating associative weights under conditions of relative safety. We would expect REM-linked recalibration to involve amygdala–vmPFC coupling changes and reduced amygdala reactivity over repeated exposure to the same threat theme. In short, REM’s neuromodulatory environment permits fear extinction–like processes and memory reconsolidation without external cost.
- NREM (slow-wave) sleep: During deep slow-wave sleep, hippocampal–neocortical replay supports memory consolidation and integration. Evidence shows replay sequences can occur in temporally compressed or even reverse order. This replay architecture provides a plausible substrate for restructuring emotional sequences offline. For QS, NREM offers a mechanism by which recent emotionally salient experiences can be integrated into broader memory networks, potentially reducing their residual charge. The result upon waking may be altered option weights (e.g., reduced avoidance bias, renewed motivation) consistent with compensatory recalibration.
- Insula and interoceptive simulation: Dreaming also engages interoceptive representations. The insula, which integrates bodily state signals, remains active during REM. Dream simulations often include visceral sensations (tightness, relief, breath changes). Upon waking, altered interoceptive tone (e.g., calmer heart rate variability or reduced sympathetic activation) may reflect overnight

adjustment. For example, after a dream involving mastery of a challenge, physiological responses to related stressors the next day may be attenuated.

- ACC and rIFG micro-arousals: The anterior cingulate cortex (ACC) and right inferior frontal gyrus (rIFG), both involved in conflict monitoring and inhibitory control, exhibit transient bursts during moments of heightened arousal in REM–NREM transitions. These micro-arousals may interrupt escalating nightmare trajectories or shift dream content. Such interventions could function as internal braking mechanisms that prevent overwhelming distress. Night terrors in children, which often terminate abruptly with awakening, may reflect an extreme version of this process. The broader point is that sleep circuitry includes control mechanisms capable of constraining emotional escalation.

5.4.4 Phenomenological Examples Consistent with the Model

Common dream experiences illustrate how emotional counterweighting could operate without invoking teleology:

- Threats turned traversable: Many individuals report recurring chase or threat dreams that gradually become less terrifying over time. Each iteration may move closer to mastery or escape. The trajectory often shifts from panic to agency. This pattern aligns with reconsolidation and extinction mechanisms operating offline. Behaviorally, individuals frequently report reduced avoidance the following day.
- Grief dreams (reunion and release): After bereavement, dreams often feature reunions or conversations with the deceased. Reports commonly describe a sense of closure, forgiveness, or emotional completion. The model interprets these as emotionally integrative episodes that reduce unresolved ledger entries, facilitating adaptive adjustment upon waking.
- Rehearsal and relief dreams: Before major stressors (exams, performances, confrontations), dreams frequently simulate failure scenarios. Although distressing in-dream, these experiences may function as low-cost rehearsal exposures. Upon waking, individuals often report diminished anticipatory anxiety. The next-day performance context may feel more manageable because the emotional spike was partially metabolized offline.

None of these phenomena require purpose-driven interpretation. They are consistent with sleep-dependent memory updating and affect regulation operating within a constraint framework.

5.4.5 Laboratory Tests

We can rigorously test the dream counterweight hypothesis with controlled experiments. Four example designs are outlined below, each with predicted outcomes. All studies would use polysomnography (to stage sleep and quantify REM/NREM), structured dream sampling (e.g., awakenings during REM for reports), and next-day behavioral assays to assess whether dreams influenced actions or physiological responses.

A. Day–Night Inversion Study (Within-Subject)

Design: Systematically induce mild, ethically approved daytime perturbations and test for compensatory inversion at night. On some days, induce a mild negative ΔL_{day} (e.g., via social stress, unsolvable tasks, mild performance failure); on other days, induce a mild positive ΔL_{day} (e.g., unexpected reward or success). Subjects then sleep in the lab. During the night, collect dream affect ratings (via REM awakenings and morning reports) and physiological measures (heart rate, EEG theta power, REM density). The next morning, present tasks offering “repair” options (e.g., helping someone, correcting a mistake, reconciliation choices) and measure selection bias.

Prediction: Dream affect D will correlate negatively with ΔL_{day} . After a negative day ($\Delta L_{\text{day}} < 0$), mean D should shift positive relative to baseline; after a positive day, D should shift negative or neutral. The inversion slope should scale with horizon H: under experimentally shortened horizons (e.g., deadline framing or older vs. younger comparison, pre-registered), the magnitude of the $D \sim -a(H)\Delta L_{\text{day}}$ slope should increase. REM intensity markers (REM density, theta power, PGO-related indices where measurable) should correlate with the magnitude of inversion.

Outcome measures: Significant negative regression of D on ΔL_{day} , stronger inversion under short-horizon conditions, and next-day behavioral tilt toward compensatory actions following stronger inversion dreams.

Null/fail condition: No reliable negative coupling between ΔL_{day} and D after controlling for circadian effects and baseline mood, and no Φ - or horizon-related scaling of inversion magnitude.

B. Targeted Memory Reactivation (TMR)

Design: During wake, associate a neutral cue (e.g., odor or tone) with either a negative perturbation (e.g., mild stressor) or a positive experience. During subsequent sleep (REM or NREM as pre-registered), present the cue to bias reactivation. Collect dream reports and next-day behavior.

Prediction: Cued memories should reappear in dreams. Under QS, the reactivated content should shift directionally toward compensation: negative memories replay with reduced threat or increased mastery; overly positive indulgent memories replay with corrective elements. Next-day choices should reflect this rebalancing (e.g., increased approach toward reparative action or moderated indulgence).

Outcome measures: Increased frequency of cued memory incorporation in dreams; systematic affective transformation of that content; next-day behavioral shift consistent with compensability logic; neural correlates in vmPFC/ACC consistent with altered valuation of the reactivated stimulus.

Null/fail condition: Cue incorporation occurs without affective shift, or dream content changes do not predict next-day compensatory behavior beyond baseline learning effects.

C. REM Suppression/Rebound

Design: Use a split-night or pharmacological REM-suppression protocol (ethically approved; e.g., REM interruption or selective suppression). Compare nights with REM suppression to normal sleep within subjects.

Prediction: If REM supports counterweight processing, suppression should attenuate dream inversion (weaker $D \sim -\Delta L_{day}$ slope) and reduce next-day compensatory bias. Rebound REM should show stronger-than-baseline inversion and stronger next-day compensatory tendencies.

Outcome measures: Reduced ΔL_{day} -D coupling under REM suppression; amplified coupling during rebound; measurable downstream behavioral differences (e.g., lower probability of selecting compensatory options after suppressed REM relative to normal REM).

Null/fail condition: REM manipulation does not alter inversion slope or next-day compensatory behavior beyond general fatigue or sleep fragmentation effects.

D. Congestion-Free vs. Congested Days

Design: Manipulate or measure waking congestion (high-demand vs. low-demand days). On high-congestion days, subjects have limited opportunity to address stressors; on low-congestion days, relief channels are accessible.

Prediction: When daytime compensatory channels are blocked (high congestion), dream counterweights should carry more of the compensatory load (stronger $D \sim -\Delta L_{day}$ coupling). On low-congestion days, dream inversion should weaken because waking

repair occurred. Next-day behavioral carryover from dreams should be larger following high-congestion nights.

Outcome measures: Significant interaction between daytime congestion and dream inversion slope; greater next-day compensatory behavior following high-congestion inversion nights; moderation of this effect by shared-resource penalty indices.

Null/fail condition: Dream inversion magnitude does not differ by congestion level and does not interact with next-day compensatory behavior.

5.4.6 Naturalistic Telemetry Signatures

Beyond the lab, if dreams truly serve as QS counterweights, we should observe structured patterns in everyday life data. Using wearable sensors, ecological momentary assessment (EMA), and structured dream diaries (“telemetry”), we can test the following predictions:

- Sleep quality moderates next-day tilt: Individuals with stronger REM density, REM continuity, or consolidated SWS on a given night should show a larger next-morning behavioral shift toward reparative or balance-restoring actions, conditional on prior-day ledger drift ΔL_{day} . For example, after a negative day, those with higher REM intensity should be more likely to initiate relief- or repair-oriented behaviors (e.g., apology, help-seeking, task completion) than individuals with comparable ΔL_{day} but lower REM metrics. Crucially, analyses must control for baseline mood, total sleep duration, and trait affect to isolate sleep-stage-specific effects rather than generic rest effects.
- Dream–behavior coupling: When dream reports contain compensatory themes (e.g., relief, mastery, reconciliation), measurable next-day behavioral differences should follow. After dreams of successful coping or resolution, individuals should display reduced avoidance and increased proactive behavior in analogous contexts the next day. This coupling must persist after controlling for baseline affect, sleep duration, and prior-day arousal. The key test is not whether people feel better, but whether dream content predicts structured behavioral adjustment consistent with compensability logic.
- Horizon sensitivity in dreams: Individuals with objectively shorter horizons (e.g., advanced age, terminal illness, externally constrained timelines) should exhibit stronger dream counterweight effects and faster morning pivot toward high- Φ actions, conditional on intact cognitive and physiological channels. That is, the slope of $D \sim -a(H)\Delta L_{\text{day}}$ should increase as H decreases. This prediction is

directional and interaction-based: horizon status alone should not predict dream valence, but horizon should amplify compensatory inversion magnitude.

- Ledger-dependent asymmetry: Dream counterweights should scale with magnitude of imbalance $|\Delta L_{\text{day}}|$, not merely with stress intensity. If dreams are QS-mediated, stronger daytime imbalance should produce proportionally stronger compensatory dream shifts, rather than a flat or threshold effect.

Operational requirements:

- Pre-register inversion slopes ($D \sim -\Delta L_{\text{day}}$) and horizon interactions.
- Include baseline mood, trait affect, and total sleep time as nuisance regressors.
- Use mixed-effects models with subject-level random slopes to capture within-person dynamics.
- Test next-day behavior shifts as mediated by dream affect rather than by general sleep quality alone.

Null/fail conditions:

- No reliable negative coupling between ΔL_{day} and dream affect D after controlling for circadian rhythm and baseline affect.
- No $\Delta L_{\text{day}} \times H$ interaction in dream inversion magnitude.
- Dream content does not predict next-day compensatory behavior beyond baseline mood or sleep duration.
- Strong REM metrics correlate with general mood improvement but not specifically with compensability-aligned actions.

5.4.7 Clinical Extensions

The counterweight framework for dreams generates clinically testable hypotheses across several conditions. These extensions are exploratory and must be evaluated with appropriate controls; they do not assume that dreams are sufficient for recovery, only that they may participate in compensatory regulation under certain neurobiological constraints.

- Post-traumatic stress disorder (PTSD) and nightmares: In PTSD, nightmares are often recurrent and distressing. From a QS perspective, nightmares may represent failed or incomplete counterweights. Elevated noradrenergic tone during REM, a well-replicated finding in PTSD, can impair the brain's ability to safely recontextualize traumatic memory traces. If REM neurochemistry prevents flexible updating (e.g., persistent hyperarousal or impaired vmPFC regulation of amygdala activity), dream content may repeatedly re-expose the individual

without achieving emotional integration. Interventions that lower noradrenergic tone (e.g., prazosin in appropriate cases) or structured scripting approaches such as Imagery Rehearsal Therapy may restore conditions under which dream content can shift from re-traumatizing replay toward compensatory processing. A clear empirical prediction is that successful treatment should increase dream variability and the presence of mastery or resolution themes, accompanied by improved vmPFC–amygdala functional coupling during REM. If symptom reduction occurs without any change in dream tone, diversity, or REM-related regulatory markers, the counterweight hypothesis would be weakened in this domain.

- Depression and anhedonia: Major depression is associated with disrupted sleep architecture, including altered REM latency, REM density, and reduced positive dream affect in many patients. The counterweight model predicts that insufficient dream compensation may contribute to persistently negative next-day tilt. If REM is fragmented or emotionally blunted, the offline opportunity to rebalance ΔL_{day} may be diminished. Treatments that improve sleep continuity and affective range—whether behavioral (e.g., behavioral activation), psychotherapeutic, or pharmacologic—should, under this model, be associated with increases in emotional diversity and compensatory themes in dream reports. A testable prediction is that increases in dream affect variability or mastery content should precede or parallel measurable improvements in next-day mood or decision tilt. If mood improves while dream content remains flat and non-diverse under careful measurement, then dream-based compensation may not be contributing meaningfully in that case.
- End-of-life (EOL) care: Hospice and palliative care clinicians frequently report vivid dreams, reconciliatory imagery, or symbolic closure themes near the end of life. The QS model predicts that as $H \rightarrow 0$, compensatory weighting intensifies, and dreams may preferentially simulate unfinished relational or existential material in forms that reduce unresolved ΔL . This does not imply that all patients will experience consoling dreams; channel integrity (sleep quality, neurochemistry, medication effects) constrains the process. A measurable prediction is that dream frequency and emotional resolution themes increase as objective or subjective horizon shortens, controlling for medication load and delirium risk. If no systematic shift in dream tone or compensatory content is observed in well-measured EOL cohorts with intact REM physiology, the horizon-scaling component would be challenged.

5.4.8 Lucid Dreams and Agency

Lucid dreaming, in which the dreamer becomes aware that they are dreaming and can exert some degree of control, provides a constrained test of the QS mechanism. If QS operates as a compensatory weighting process rather than a rigid script, then lucidity should modulate counterweight effects depending on how agency is used.

Specifically, if a lucid dreamer intentionally steers the dream toward compensatory or mastery themes (a high- Φ scenario), we predict amplified next-day compensatory tilt relative to non-lucid REM of similar baseline affect. In contrast, if lucidity is used for indulgent or escapist themes unrelated to repair (a low- Φ scenario), the counterweight signal should be attenuated or redirected, and next-day tilt should reflect reduced compensatory shift.

An experimental design could randomize lucid dreamers to two conditions upon becoming lucid: (1) a “repair” prompt (engage a feared scenario, rehearse reconciliation, complete an unfinished task), or (2) a neutral or indulgent prompt (engage in non-compensatory exploration). Primary outcomes would include next-day mood shift, behavioral choice bias toward high- Φ options, and QS-related neural markers during subsequent decision tasks. A falsifier would be no systematic difference in next-day compensatory tilt between repair-directed and indulgence-directed lucid dreams under adequate power and control of expectancy effects.

Taken together, these clinical and agency-based extensions illustrate how the dream counterweight hypothesis can be tested without invoking teleology. If compensatory dream mechanisms are real, they should scale with neurochemical conditions, horizon, and intentional use. If those signatures fail to appear under rigorous measurement, the counterweight account must be revised or rejected.

5.4.9 Distinguishing the QS Account from Rival Theories

There are several established theories of dreaming and emotional regulation. To evaluate the counterweight hypothesis rigorously, we must identify patterns that distinguish the Queue System (QS) account from these alternatives. Three primary rival frameworks are considered here.

- Homeostatic rebound: After an emotionally extreme day, affect may naturally swing in the opposite direction through standard homeostatic or opponent processes. On this account, if you are stressed, the next day may feel relatively calmer simply because physiological systems reset. REM sleep itself may produce a generic rebound of whatever state was underrepresented during

waking (e.g., more positive imagery after a painful day, more threat imagery after an unusually pleasant day). This could explain day-night affect inversions. However, QS predicts something more specific: dream content should track the content of the imbalance, not merely its sign. For example, if a person experienced a humiliation event, the counterweight dream should disproportionately involve mastery or social repair themes, not simply generic positive mood. Homeostatic rebound does not inherently predict content-selective inversions tied to the structure of the day's ledger, nor does it predict stronger inversion slopes when the horizon is short.

- Threat simulation theory (TST): One influential evolutionary account holds that dreams simulate threats so that individuals can rehearse defensive responses. TST predicts frequent threat content and may explain why many dreams are negative or anxiety-driven. However, TST primarily addresses rehearsal of danger scenarios and does not posit symmetric counterweights for positive excess or relief states. QS, by contrast, predicts bidirectional counterbalancing: negative days should yield compensatory positive or mastery dreams, and unusually positive days may yield cautionary or stabilizing content. In addition, QS predicts horizon scaling, meaning counterweights intensify when perceived time is limited. Threat simulation alone does not predict systematic strengthening of dream inversion effects as life horizon shortens.
- Standard memory consolidation: A widely accepted view is that dreams reflect replay and integration of recent memories, especially emotionally salient ones. On this account, dreaming simply mirrors what was encoded during the day. However, replay alone does not imply directional correction. If consolidation were neutral, dreams should reproduce the day's dominant themes proportionally, not invert them in a structured way. QS predicts selective weighting during replay, where memory traces that restore balance are preferentially elaborated. Consolidation theories also do not predict next-day behavioral tilt toward reparative actions, whereas QS predicts that compensatory dreams should measurably shift subsequent choices.

A clear discriminator across these frameworks is quantitative structure. Under QS, we expect a content-sensitive inversion pattern such that dream affect D satisfies approximately:

$$E[D | \Delta L_{\text{day}}, H] \approx -a(H) \Delta L_{\text{day}} + \eta, \text{ with } a(H) \text{ increasing as } H \text{ decreases.}$$

That is, dream affect should scale negatively with the prior day's ledger change, and the magnitude of this inversion should grow as horizon shortens. In addition, QS predicts

content matching (repair themes following relational harm, mastery themes following failure, relief themes following deprivation), not merely general mood shifts.

Rival accounts may explain portions of the data. For example, REM after stress or sleep deprivation can alter emotional tone, and homeostatic regulation may contribute to baseline stabilization. The QS claim is narrower and more demanding: dream inversion should be (i) proportional to the magnitude of ledger imbalance, (ii) sensitive to horizon, and (iii) predictive of next-day behavioral shifts toward compensation.

To rule out confounds, analyses must control for total sleep time, REM density, sleep fragmentation, circadian phase, baseline mood, and trait affect. If, after such controls, the Φ -linked dream inversion slope remains and scales with H , that supports QS. If inversion reduces to generic rebound, circadian fluctuation, or simple replay proportional to day content, then QS adds no explanatory value.

Null/fail condition: If dream affect correlates with prior-day affect only through generic valence rebound (independent of content specificity and independent of horizon), and if including Φ - and H -based terms does not improve prediction beyond standard homeostatic or consolidation models, then the QS account of dream counterweights is not supported.

The distinguishing burden is therefore precision. QS survives only if it predicts patterns that rival theories cannot reproduce without explicitly incorporating a fairness-like constraint. If the data show only global mood rebound or simple replay, the QS interpretation should be rejected.

5.4.10 Fail Conditions for the Dream Counterweight Hypothesis

What evidence would disprove the idea that dreams serve as low-cost counterweights for QS? We identify several clear, preregisterable fail conditions:

- No reliable day-night affect inversion: If large, well-controlled studies consistently find no negative relationship between daytime ledger change (ΔL_{day}) and dream affect (D), the counterweight hypothesis weakens. For example, if after statistically worse days (as defined by prespecified HCl-based thresholds) dreams are no more likely to be positive—or are random with respect to ΔL_{day} —then dreams are not functioning as inverse adjustments.
- No horizon modulation of dream effects: If individuals nearing the end of a feasible horizon ($H \rightarrow 0$) do not show stronger or more frequent counterbalancing dream patterns, QS loses predictive specificity. Operationally: after controlling for age, illness burden, medication, and REM architecture, dream affect magnitude

should not systematically differ as H shrinks. If terminal and non-terminal participants with matched daytime ledger shocks exhibit indistinguishable dream responses, horizon-sensitive modulation is unsupported.

- REM manipulation does not bias content: A stronger experimental falsifier: if targeted memory reactivation (TMR), REM extension/suppression, or REM fragmentation systematically alter dream content without altering next-day HCl trajectories, then dreams may reflect replay processes rather than ledger-adjusting dynamics. Conversely, if altering REM architecture produces predictable shifts in next-day affect consistent with counterweight predictions, the hypothesis survives. Null behavioral effects under controlled REM perturbation challenge the mechanism.
- REM removal without ledger drift: If REM deprivation (sleep restriction, pharmacological suppression) eliminates emotional dreams yet leaves next-day affect variance and drift statistically unchanged, dreams are unlikely to be necessary components of QS adjustment. For example, if PTSD patients on prazosin show reduced nightmares without measurable change in subsequent ledger variance or mean reversion patterns, compensation may occur via alternative pathways.
- Rival mechanistic models outperform QS residuals: If dream affect is fully explained by known homeostatic, threat-simulation, memory consolidation, or arousal-regulation mechanisms—without requiring horizon- or ledger-dependent terms—then the counterweight interpretation loses unique explanatory value. Formally: if a non-QS model M_0 explains dream affect with equal or better out-of-sample predictive performance than a QS-residual model M_1 that includes $\Phi(H, L)$, then the added QS structure is unjustified.

If any of the above findings were robustly demonstrated—especially in preregistered studies with transparent coding and adequate power—we would concede that dreams, while psychologically meaningful, are not systematically contributing to ledger stabilization.

5.4.11 Ethics and Measurement Notes

Given the sensitivity of dream research, ethical and methodological rigor are non-negotiable.

- Participant welfare: Nightmare induction or emotionally provocative paradigms must avoid intentional harm. Brief awakenings and minimal-disturbance protocols are required. No study should deliberately induce daytime trauma for measurement purposes. Clinical populations (e.g., hospice patients, PTSD

patients) require enhanced consent, monitoring, and debriefing procedures. Data collection never supersedes participant well-being.

- Measurement of dream affect: Dream valence should be assessed using multi-channel approaches. Immediate post-REM self-report scales (valence, arousal, thematic intensity) should be combined with physiological markers (heart rate variability, skin conductance, REM density, EOG spectral features). Because recall degrades rapidly, awakenings should be standardized in timing and frequency. Coding categories must be preregistered and independent raters blinded to daytime ledger state. Inter-rater reliability thresholds (e.g., Cohen's $\kappa \geq 0.60$) should be prespecified.
- Operationalizing ΔL_{day} : Daytime ledger change should be estimated using the same HCI framework defined earlier in the book, including pre-specified smoothing windows and variance normalization. Thresholds for "bad days" or "good days" must be defined before analysis to avoid post hoc categorization.
- Transparency and reproducibility: All analytic decisions—including horizon proxies, REM scoring criteria, coding definitions, and statistical models—must be preregistered. Raw dream transcripts (anonymized), physiological time series, and analysis code should be made available when ethically permissible. Analysis windows must be fixed in advance to prevent p-hacking. Replication across labs is required before strong claims are made.

Takeaway: Dreams may function as a night-shift mechanism—an off-line, low-cost, plasticity-rich domain in which the emotional system adjusts variance accumulated during the day. Under the Law of Fairness, we expect horizon-sensitive inversions between daytime ledger shocks and dream affect. If such inversions fail to appear under rigorous testing, the counterweight hypothesis should be rejected.

5.4.12 Where we go next:

To separate QS-specific structure from ordinary sleep dynamics, the next section introduces a residual framework. After modeling utility, arousal, conflict, risk, and timing, we test whether a feasibility-weighted control term remains—precisely the variance-sensitive residual predicted by QS.

5.5 Research Notes: QS-Residuals After Nuisance Modeling

This section provides a methods blueprint for identifying QS-residuals, defined as signals in behavior or neural activity that remain after accounting for standard decision drivers and shared-resource effects. The objective is not to produce a single decisive regression that proves QS, but to isolate a reproducible residual pattern that survives model competition, cross-validation, and specification stress tests. The task is to determine whether a feasibility-weighted signal consistent with QS remains once conventional predictors have been fully modeled.

5.5.1 Identification Recipe (Bird's-Eye View)

To isolate QS-residual structure, we proceed in the following sequence.

Define preregistered causal models. Formally specify the assumed causal structure using diagrams or structural equations and preregister this structure. For individual utilities u , assume direct effects on choices. Conflict and arousal influence control signals. The running ledger L and horizon H jointly determine compensability metrics Φ . External resource context R contributes shared penalties through shadow prices $\lambda r(t)$. In shorthand:

- ledger L and horizon $H \rightarrow \Phi$
- congestion $R \rightarrow \lambda r \rightarrow$ social penalty
- utilities \rightarrow choices
- conflict and arousal \rightarrow control signals

Encoding these relationships clarifies where QS terms enter, namely downstream of L and H . The directed model determines which variables are controls and which are candidate QS predictors. All modeling decisions are specified before inspecting QS coefficients.

Fit a comprehensive nuisance backbone model. Construct the strongest feasible predictive model of choices and neural responses without QS-specific structure. This backbone must incorporate all established drivers that plausibly explain behavior.

Utility and immediate reward variables, such as option value and expected return.

Conflict or cost measures, including entropy of the choice distribution, reaction time adjustments, or Stroop-like interference indices.

Risk and ambiguity parameters, including magnitude, variance, outcome uncertainty, and ambiguity sensitivity.

Arousal and stress indicators such as skin conductance, heart rate variability, pupil diameter, or cortisol proxies where available.

Habit or fatigue factors, including trial number, session time, prior choices, and time-on-task trends.

Learning signals if applicable, such as reward prediction errors, reinforcement learning state values, or eligibility traces.

Session and person-level covariates, including age, time of day, sleep quality, medication status, scanner drift, and other sources of mundane variability.

Random intercepts and slopes for subjects and sessions to account for baseline heterogeneity.

This nuisance model should be as comprehensive and high-performing as possible, using regularization or Bayesian priors to avoid overfitting and nonlinear terms where theoretically justified. The goal is to explain variance attributable to standard theories before testing QS structure. Only once this backbone fits well under cross-validated log likelihood, predictive log loss, or cross-validated R^2 do we proceed.

Add QS predictors. Introduce the hypothesized QS variables and test whether they account for additional out-of-sample variance.

The feasibility-of-compensation score for each option, $\Phi(u; L, H, R)$, operationalized through measurable features such as predicted ReliefGain or HarmRisk. Φ captures how selecting option u alters the probability of remaining within the neutral closure band $|L(T)| \leq K$.

The horizon interaction, typically specified as $\Phi \times g(H)$, where $g(H)$ increases as H decreases. This may be modeled as $1/H$, a monotonic spline in H , or a short-horizon indicator. The central prediction is that the marginal effect of Φ strengthens as remaining feasible horizon shrinks.

The shared-resource penalty term $\sum r(t) \Delta r(u)$, where $\Delta r(u)$ measures resource draw and $r(t)$ reflects time-varying shadow prices. This term incorporates congestion costs. If Φ already embeds these penalties, they should not be reintroduced separately to avoid double counting.

In neural models, QS terms are added as parametric regressors aligned with task events in a GLM or analogous regression framework. In behavioral models, they enter the utility function or decision rule. The question at this stage is whether QS structure improves prediction beyond the nuisance backbone.

Evaluate incremental predictive value. Improvements must be assessed strictly out of sample using nested cross-validation or subject-level and time-based train/test splits. Metrics may include cross-validated log likelihood, predictive log loss, mean squared error for continuous outcomes, Brier score or log loss for categorical outcomes, and ROC-AUC for classification. Information criteria such as WAIC or PSIS-LOO may supplement but do not replace out-of-sample validation. The critical test is whether adding QS terms yields a statistically reliable and practically meaningful reduction in prediction error relative to the backbone model.

Examine coefficient stability and generalization. QS coefficients should retain sign and approximate magnitude across folds, sites, or subsamples. Credible or confidence intervals should exclude zero consistently in held-out data. Effect sizes should be reported with uncertainty, and partial dependence or marginal effect plots should illustrate how predicted outcomes vary as a function of Φ while holding nuisance predictors fixed.

Stress-test robustness. Re-estimate models under alternative specifications and sanity checks.

Use alternative nuisance sets to confirm that QS effects do not depend on a single modeling choice.

Test nonlinear transformations or alternative interaction forms for L, H, or Φ , including spline bases or threshold models, to ensure the effect is not an artifact of functional form.

Assess multicollinearity using variance inflation factors or posterior correlation matrices. If Φ is highly correlated with utility or arousal terms, consider orthogonalizing or residualizing Φ to verify that it carries unique signal.

Where feasible, use instrumental variable approaches or experimental manipulations to address potential endogeneity of H or L.

Include negative control trials in which Φ is constant across options. In such trials no QS effect should be detectable. A significant coefficient under these conditions indicates misspecification.

Report effect geometry, not just p-values. If QS terms hold up, results must be reported transparently with effect size estimates and uncertainty intervals. For example, adding Φ may improve predictive log loss by X percent \pm Y, or the coefficient of Φ in predicting ACC activity may be 0.5 with a 95 percent interval of [0.2, 0.8]. Visualizations such as partial dependence plots should show how predicted outcomes change as a function of Φ while

holding other predictors constant. Replication across independent subsets of data or different laboratories should be demonstrated where possible. Running analyses separately in two halves of a sample or across experiment sites provides a direct check on stability. The aim is to present a complete characterization of the QS-residual: its magnitude, where it appears, and how stable it remains across conditions and samples.

The objective of this procedure is not to confirm QS by construction but to attempt to eliminate it through rigorous nuisance control. A QS-residual that persists after these steps constitutes convergent evidence consistent with the model. If the residual vanishes under proper cross-validation and specification checks, the hypothesis does not survive.

5.5.2 The Nuisance Backbone (What Must Be Controlled)

As outlined above, the backbone model must capture all ordinary influences on decisions and neural signals before any QS-specific structure is introduced. The goal is to remove plausible alternative explanations so that any remaining effect cannot be attributed to standard drivers.

Utility and expected value. Trial-by-trial utility of each option and its immediate hedonic impact. For example, monetary value, reward points, or subjectively rated pleasure and pain. In neuroimaging, this typically enters as a parametric modulator of task events in regions associated with subjective value. If an option is highly rewarding, the model must already account for the fact that it is likely to be chosen and that canonical valuation regions will respond.

Conflict and cognitive control. Metrics that capture how difficult or internally conflicting a choice is. This may include entropy of the choice distribution, response conflict measures, Stroop-like interference indices, or reaction time adjustments for hard versus easy trials. Activity in regions associated with control signals can often be explained by such predictors. Including explicit conflict regressors prevents attributing generic control-related variance to QS.

Risk and ambiguity. Variables reflecting outcome uncertainty, such as variance, probability dispersion, ambiguity, or risk magnitude. Many individuals display risk aversion or ambiguity sensitivity, and both behavior and neural activity can scale with these parameters. Explicitly modeling these effects ensures that systematic relationships between uncertainty and choice are not misclassified as QS structure.

Arousal and stress level. Physiological and endocrine measures that reflect autonomic activation, including skin conductance, heart rate, heart rate variability, pupil diameter, and cortisol when available. Elevated arousal can alter response caution, decision

thresholds, and neural gain. Including these measures prevents mistaking state-dependent modulation for compensability dynamics.

Habit, fatigue, and sequential effects. Variables capturing autocorrelation and time-on-task dynamics, such as previous choice, trial number, session duration, or accumulated effort. These predictors account for routinization, inertia, or performance degradation over time. Without them, slow drifts in behavior could be misinterpreted as ledger-related adjustment.

Learning signals. If the task involves feedback, include standard learning model predictors such as reward prediction errors, state values, policy values, or eligibility traces. These explain systematic updating in behavior and neural responses. Excluding them risks attributing learning-driven variance to QS.

Sociodemographic and session-level covariates. Age, relevant medications, sleep quality, time of day, and other contextual factors that plausibly influence affect, arousal, or valuation. In neuroimaging, include motion parameters and scanner drift terms. In behavioral data, include experimenter or batch indicators when appropriate. These controls absorb mundane but potentially confounding variance.

Random effects. In hierarchical data, include random intercepts for subjects and sessions, and random slopes for key predictors where justified. This accounts for baseline differences in response tendencies and heterogeneity in sensitivity to risk, value, or conflict. Proper hierarchical structure prevents population-level effects from being driven by a small subset of individuals.

The nuisance backbone must be evaluated for goodness of fit before introducing QS terms. Cross-validated predictive performance should indicate that standard drivers explain the bulk of systematic variance. If QS terms appear significant in a weak or under-specified backbone, the result should be treated with skepticism. Only after the backbone demonstrates adequate out-of-sample performance do we proceed to test whether QS structure explains residual variance beyond established mechanisms.

5.5.3 QS Terms and Operational Features (What We Add)

When we augment the nuisance backbone with QS structure, we must translate the abstract variables $\Phi(u; L, H, R)$, L, and H into measurable features. These features must be defined in advance, computed without peeking at outcomes, and aligned with the theoretical role each term plays in the model.

Feasibility-of-compensation features. Conceptually, $\Phi(u; L, H, R)$ represents how selecting action u alters the probability of ending the relevant episode with a ledger

inside the neutral band $|L(T)| \leq K$. Because this probability is not directly observable, we approximate Φ using task-defined features that capture compensability.

ReliefGain. The expected immediate improvement in ledger position produced by option u . In practice this may be the expected reward, affective uplift, or reduction in ongoing negative state, expressed in the same scale used to estimate L . ReliefGain should be computed from task structure or model-based expectations, not from observed outcomes.

RepairGain. If option u corrects or mitigates a prior negative impact, the estimated reduction in previously accumulated harm enters as positive compensability. This term captures apology, restitution, or corrective actions when those are explicitly modeled.

HarmRisk. The expected increase in ledger variance or downside exposure associated with u . This enters with a negative sign. It may be operationalized as expected loss magnitude, tail risk, or probability-weighted negative outcomes. HarmRisk should reflect forward-looking exposure, not realized loss.

OptionFlex. A measure of preserved optionality. Actions that leave more future states reachable, or that avoid irreversible depletion of time or resources, increase compensability. Flexibility can be proxied by remaining choice set size, resource slack, or reversibility indicators defined by the task.

Social penalty. A shared-cost term reflecting resource draw that reduces feasibility for others. When relevant, this is implemented as a separate $\sum_r \lambda r(t) \Delta r(u)$ component rather than folded implicitly into Φ , unless the operational definition of Φ already includes it. The implementation must avoid double counting.

All Φ components must be computable from ex ante task information or model-based expectations. They should be linear or smoothly nonlinear functions of predefined features and not tuned post hoc to maximize model fit.

Horizon scaling. The influence of compensability is predicted to depend on remaining feasible horizon H . Operationally, this may be implemented as an interaction $\Phi \times g(H)$, where $g(H)$ increases as H decreases. H may be directly measured, experimentally manipulated, or proxied by remaining trials, time budget, or externally imposed deadlines. Nonlinear forms such as $1/H$ or monotonic splines are acceptable if specified in advance. The key empirical question is whether the marginal effect of Φ strengthens as H shrinks.

Shared-resource penalty. We explicitly model $\sum_r \lambda r(t) \Delta r(u)$, where $\Delta r(u)$ quantifies resource consumption and $\lambda r(t)$ reflects time-varying shadow prices. In field settings,

$\lambda r(t)$ may be derived from observable congestion metrics such as queue length, wait times, or budget utilization. In experiments, it may be manipulated by varying group size or resource scarcity. If $\lambda r(t)$ correlates with neural or behavioral responses independently of Φ , that effect must be absorbed by the penalty term rather than attributed to QS structure.

Ledger magnitude and sign. Although the simplified presentation often omits L from the regression equation, operational models may include current ledger level $L(t)$ or its sign as moderators. The influence of Φ may differ when the ledger is strongly negative versus near neutral. However, including L requires that Φ explain variance beyond a simple imbalance effect. If Φ adds no explanatory power once L is included, it is not carrying unique signal.

Offline counterweight activity. When relevant data are available, lagged measures of off-line adjustment may be incorporated. For example, previous-night dream affect D can be entered as a predictor of next-day decision tendencies, consistent with the counterweight hypothesis developed earlier. Such features must be time-aligned and specified prior to analysis. Mediation analyses, when used, should test whether off-line variables account for links between prior-day ledger change and subsequent choices without circularity.

All QS features must be aligned correctly in time, whether concurrent or lagged, and defined before analysis. No post hoc feature construction is permitted. If a feature fails to predict, it remains part of the preregistered model and is reported as null. Operationalization should be locked by theory and task design rather than adapted to improve fit after inspecting results.

5.5.4 Canonical Model Forms (Neural and Behavioral)

To make this concrete, here is how models with QS terms typically look for neural data (fMRI or EEG) and for behavioral data (choices and response times). These are general forms rather than rigid equations.

Neural model (ROI or voxel-wise GLM): We model the neural response $y_{ROI}(t)$ (for example, the BOLD signal in a specific brain region or a single voxel at time t) as:

$$y_{ROI}(t) = \text{Backbone}(t) + \gamma_\Phi \Phi(u(t); L(t), H(t), R(t)) + \gamma_H [\Phi(u(t)) \times g(H(t))] + \gamma_S \sum \lambda r(t) \Delta r(u(t)) + \varepsilon(t)$$

$\text{Backbone}(t)$ represents the combined effect of all nuisance regressors at time t (task events, utilities, conflict, risk, learning, arousal, etc., convolved with the hemodynamic response for fMRI where appropriate). The QS regressors have coefficients γ_Φ , γ_H , and

γ_S respectively: the compensability value Φ of the option being evaluated or chosen at time t ; the interaction of Φ with a horizon-scaling function $g(H)$ that increases as remaining horizon shrinks (for example $g(H) = 1/H$ or a prespecified monotonic spline); and the shared resource load term $\sum \lambda r(t) \Delta r(u(t))$. $\varepsilon(t)$ is residual noise after accounting for temporal autocorrelation and model structure.

In a mass-univariate fMRI analysis, this model is estimated separately for each voxel or ROI. We would predict, for example, a positive γ_Φ in value-sensitive regions such as vmPFC or OFC if more compensatory options elicit stronger valuation signals, conditional on the backbone. Regions associated with inhibitory control or threat processing may show the opposite pattern if low- Φ or high-harm-risk options recruit braking or caution mechanisms. A positive γ_H would indicate that neural sensitivity to compensability increases as horizon shrinks. γ_S may appear in regions implicated in conflict or salience processing when shared resource load is high. Crucially, inference depends not only on coefficient sign but on whether adding these regressors improves out-of-sample model fit and whether posterior or confidence intervals exclude zero.

Behavioral model (choices): For discrete choices, QS terms enter the decision policy. A standard formulation uses a softmax mapping from latent utilities to choice probabilities:

$$P(u \text{ chosen at } t) = \exp(U(u,t)) / \sum k \exp(U(k,t))$$

with

$$U(u,t) = \theta_0 + \theta_U U_{\text{standard}}(u,t) + \theta_C C(u,t) + \theta_\Phi \Phi(u; L(t), H(t), R(t)) + \theta_H [\Phi(u) \times g(H(t))] + \theta_S \sum r(t) \Delta r(u,t)$$

Here U_{standard} captures ordinary utility terms such as expected value, and C captures cost or control-related predictors. $\Phi(u; L, H, R)$ is the QS compensability term. In a two-option case, this reduces to a logistic regression on the difference in utilities. A positive θ_Φ indicates that, controlling for standard utility and costs, higher compensability increases the probability of selection. Horizon scaling can be implemented through the interaction term or through prespecified short- versus long-horizon contrasts.

Behavioral model (reaction times): For continuous measures such as response time, a common approach is to model $\log(RT)$, since RTs are positive and typically skewed:

$$\log(RT(t)) = \text{Backbone}(t) + \eta_\Phi \Phi(u(t)) + \eta_H [\Phi(u(t)) \times g(H(t))] + \eta_S \sum r(t) \Delta r(u(t)) + \varepsilon(t)$$

$\text{Backbone}(t)$ includes standard predictors of latency such as difficulty, conflict, and arousal. The signs of η_Φ and η_H depend on coding conventions. For example, if higher Φ represents more compensatory options, a negative η_Φ would indicate faster responses

for more compensatory actions, whereas a positive ηH on the interaction term would indicate that low- Φ options incur additional delay when horizon is short. Interpretation must be anchored to the specific coding of Φ and $g(H)$.

Hierarchical modeling: In many datasets, subjects differ in baseline tendencies and in sensitivity to QS terms. A hierarchical model allows subject-level random intercepts and random slopes on Φ and related interactions, for example:

$$\gamma_{\Phi,i} \sim \text{Normal}(\gamma_{\Phi,\text{group}}, \sigma_{\Phi})$$

Model comparison should be performed using cross-validation or predictive criteria applied consistently to backbone and QS-augmented models. Hierarchical structure allows partial pooling across participants and conditions while guarding against overfitting and spurious individual effects.

In summary, these model forms are straightforward extensions of standard GLMs and logistic or linear models. The novelty lies not in mathematical complexity but in the interpretation of the added regressors. If QS is real, the compensability and horizon-scaled terms should explain residual variance beyond established drivers, and they should do so robustly under out-of-sample evaluation.

5.5.5 Cross-Validation and Out-of-Sample Checks

Given the complexity of these models, confirming generalizability is paramount. We list several best practices for cross-validation (CV) and out-of-sample testing to ensure any QS effects are not overfitting artifacts or flukes:

- Blocked or grouped CV: Instead of randomly splitting trials, we hold out whole blocks of data that have logical cohesion. For example, entire days or entire subjects might be held out as test sets. This way, the model must predict new situations or new individuals it has not seen, which is a much tougher test than interpolating within the same person's data. If QS truly captures something fundamental, it should help predict a new person's behavior after training on others, or predict later trials after training on earlier trials. It also prevents leakage of information, since adjacent trials often share context and random CV can overestimate performance.
- Permutation tests for significance: To validate that QS improvements are not just due to chance configurations, we can do label shuffling. For instance, shuffle the Φ values across trials within sensible blocks that preserve some structure. This breaks any true link between Φ and outcomes. The model with shuffled Φ should show no improvement over baseline on average. If our real data improvement is

larger than 95 percent of shuffle cases, that is evidence the effect is real. Similarly, we can shuffle horizon labels or λ values to see if any structure remains.

- Temporal generalization tests: Train the model on the first portion of each session, when horizons were relatively long and people were fresh, and then test on the later portion, when horizons are shorter and fatigue may set in. QS predicts certain effects will grow over time as horizon shortens. A model that includes time varying QS terms should predict later behavior better than one without. If we see the QS model's performance diverge positively in later blocks, or as a function of horizon, that is a good sign.
- Multi site or multi cohort replication: If possible, collect a second dataset in a different lab or with a demographically different sample. Fit the model on one dataset, then test on the other, possibly doing a small adaptation of intercepts if needed. Does the QS effect translate? Alternatively, fit both separately and do a meta analysis of the QS coefficients. Do they agree in direction and magnitude? True effects should not depend on one specific sample.
- Report key metrics of improvement: To convey results, we will report things like $\Delta\text{elpd}_{\text{loo}}$ (PSIS-LOO) or ΔWAIC , meaning the change in information criterion when adding QS terms, improvements in classification accuracy or ROC-AUC for choice prediction, reductions in log loss for probabilistic predictions, and increases in variance explained, such as partial R^2 , for neural signals. For example, we might say “Adding QS terms increased variance explained in vmPFC from 20 percent to 30 percent” or “It reduced combined AIC by 15 points.” We should be cautious: even a small improvement can be meaningful if it consistently appears across validation folds and sites. Conversely, a large seeming improvement that fails to replicate is suspect.

In summary, we do not fit the model once and declare victory. We test it like an opponent on new data and difficult splits. QS, as a proposed law-level mechanism, needs to prove itself by predicting the unseen.

5.5.6 Collinearity, Instruments, and Causal Leverage.

While regression models are useful, we must disentangle QS effects from tightly correlated factors. Also, we want evidence for causality, not just correlation. A few strategies:

- Collinearity checks: We will quantify how correlated the QS predictors are with other predictors. If variance inflation factors for Φ or $\Phi \times g(H)$ are very high, for example greater than 5 or 10, that indicates effects might not be distinguishable from some linear combination of nuisance factors. For instance, Φ may be

correlated with utility if actions that are very good for the ledger also tend to be high immediate reward. To address this, one approach is orthogonalization: regress Φ on utility and use the residual component of Φ not explained by utility as the predictor. This ensures we are testing the unique contribution of compensability beyond immediate reward. Alternatively, if theory permits, we design tasks where Φ and utility are experimentally independent, for example constructing option sets where some high-utility options have low Φ and vice versa. We should also inspect posterior correlation matrices in Bayesian fits. If the correlation between the coefficient on utility and the coefficient on Φ is near 1 or -1, the model cannot reliably distinguish them. In that case, we report this and interpret cautiously.

- Instrumental variables for Φ and H : Ideally, we manipulate the key independent variables. For horizon H , we can impose deadlines or time limits in experiments to create exogenous variation in remaining time, such as telling participants a task will end unexpectedly soon versus allowing them to believe they have a longer session. For shared-resource load $\lambda(t)$, we can experimentally vary congestion, for example by simulating many people needing help at once, other times not. These serve as instruments that change Φ or λ without directly altering baseline utility. A related example is priming perceived time horizon by asking participants to think about mortality versus a neutral topic, noting that this may also shift arousal or valuation and therefore must be treated as a horizon manipulation with potential confounds rather than a clean instrument. Similarly, sleep interventions such as REM manipulation can serve as instruments for dream-driven Φ components: suppressing REM should specifically reduce the dream contribution to compensability without directly altering task utility. Using these manipulations, we can conduct analyses such as two-stage least squares or causal mediation to verify that when Φ or H is shifted exogenously, outcomes change accordingly. This provides stronger causal evidence than correlation alone.
- Mediation tests (dreams as mediator): One specific causal pathway QS proposes is: daytime imbalance ΔL_{day} influences next-day compensatory tilt. If dreams are truly implementing QS off-line, then some of the effect of a bad day on next-morning behavior should be mediated by dream affect D. We can test this by measuring ΔL_{day} , dream affect D, and next-day behavior. Do we see that ΔL_{day} predicts next-day repair behavior less strongly when controlling for D? A formal mediation analysis or cross-lagged panel model could quantify this. For example, bad days may lead to more nightmares $\Delta L_{\text{day}} \rightarrow D$, and nightmares may lead to more “making amends” behavior the next day $D \rightarrow$ next-day choice, such that the direct ΔL_{day} to next-day choice link weakens or vanishes when accounting for D.

This would suggest that dreams carry part of the causal load of QS. Careful controls are required: perhaps people who had bad days also slept poorly overall. We must therefore control for total sleep time, sleep quality, and general fatigue to isolate the dream-specific pathway.

By addressing collinearity and incorporating instrumental and mediation strategies, we reduce the risk that QS signals are simply reflections of utility, arousal, or learning. If QS dynamics involving horizon, dreams, and shared load are operative, their influence should persist under these stricter causal probes rather than collapsing into standard predictors.

5.5.7 Negative Controls and Falsification Handles

We have stressed building evidence for QS, but equally important is designing checks that would falsify QS if it is false. Negative controls are scenarios where, by design, QS should have nothing to operate on, so any detected QS effect indicates model misspecification.

- Content-irrelevant trials: Include trials in which compensability is constant across options. For example, in a task with long-term impact and occasional neutral trials, we can insert catch trials where all options are equivalent in terms of long-term ledger impact $\Phi \approx 0$. QS predicts no differential effect on these trials. If regression coefficients on Φ remain non-zero on such trials, or if shuffling which trials are labeled compensatory does not eliminate the effect, then Φ may be capturing another correlated factor rather than compensability.
- Sham congestion cues: Introduce visual or contextual cues suggesting congestion without actual shared-resource consequences. For example, display a “system is busy” message that does not affect real payoff or resource allocation. If participants respond to this sham cue as if it were real congestion, then observed λ effects may reflect generic arousal or caution rather than QS. Proper negative control requires informing participants afterward that the congestion cue was fake to ensure no real resource constraints were present. The expected outcome under QS is a null effect in such sham conditions.
- Placebo horizons: Provide misleading statements about time horizon, such as telling participants “this task will end soon” when in fact it will not. After verifying perceived horizon through questionnaires, we can test whether behavior shifts with perceived horizon alone. QS predicts that compensability effects should depend on perceived feasible horizon rather than the experimenter’s schedule. If no change in perceived horizon occurs, no QS effect should appear. If perceived

horizon changes but behavior does not, that challenges the horizon component of QS.

If these careful falsification handles still yield QS-consistent effects, confidence increases that the signals are not artifacts of arousal, confusion, or experimental design. If QS effects appear even in clearly irrelevant or placebo conditions, that is a red flag suggesting misspecification.

5.5.8 Multi-Agent Analyses (Co-Movement and Priority)

QS is not only about individual psychology but about coordination across individuals who share constraints. To test this systems-level implication, we extend analysis beyond single-person data.

- Menu co-movement across individuals: Suppose multiple people share a resource pool, such as family members sharing caregiving time or patients in a hospital ward sharing medical staff. QS predicts that when shared resource availability $R(t)$ fluctuates and $\lambda(t)$ rises or falls, individual menus shift in a coordinated way. One test is to compute, for each person, an index of low-draw option selection and examine cross-person correlations in that index across time. If QS operates at a shared-resource level, changes in $R(t)$ should induce synchronized adjustments beyond what would be expected from random alignment. Controls for common external events such as weather or news are necessary to avoid spurious correlation.
- Horizon-priority effects in groups: When multiple people draw on the same channel, those with shorter horizons should reduce usage less and retain priority access to high-value options, while those with longer horizons should voluntarily shift toward lower-intensity substitutes. For example, in a hospital setting, patients closer to end-of-life may receive high-intensity treatments, while those with longer expected horizons accept lower-intensity care under scarcity. In experimental settings, a multiplayer task with limited shared supplies can test whether individuals with experimentally shortened horizons reduce resource usage less than long-horizon participants. If all individuals reduce usage equally or long-horizon individuals continue to consume heavily under scarcity, the systems-level QS prediction is weakened.

Null results in these multi-agent analyses would challenge the claim that QS operates as a global fairness constraint. If no coordinated adjustment or priority structure appears under shared scarcity, the theory loses real-world plausibility at the systems level.

5.5.9 Simulation-Based Calibration and Power

Before and during data collection, we need to ensure our methods can actually detect QS effects if they are present, and that our sample sizes are sufficient.

- Generative simulations: First, construct simulated datasets with known parameters, some with QS active and some without. For instance, simulate choice and neural responses for 100 agents in a task where QS is programmed into the process, such as avoiding low- Φ options as horizon shrinks. Also simulate the same scenario where agents make decisions with only ordinary reward and conflict dynamics and no QS mechanism. Then run the full analysis pipeline, including nuisance regressors, cross-validation, and stress tests, on these datasets. We check two things. First, sensitivity: if QS is truly in the simulation, does our pipeline recover a significant QS effect and the related coefficients? Second, specificity: if QS is absent, does our pipeline correctly find no QS effect, meaning it does not hallucinate patterns from noise? Ideally, tune the pipeline parameters, such as prior widths, feature scaling, and cross-validation scheme, until it reliably detects QS when present and yields null results when absent. This guards against both Type I and Type II errors. We may find, for example, that if QS effects are small, we need a certain number of trials or certain design tweaks to detect them. These simulation-based checks are the unit tests for our experimental and analysis plan.
- Power analysis: Based on either pilot data or simulations, we estimate how large the sample needs to be to have a high probability of confirming a QS effect if it truly exists. Because the analyses are complex and can involve hierarchical models, traditional closed-form power formulas may not apply. Instead, use bootstrap procedures or simulation-based power calculations. We target 90 percent power to detect a moderate QS effect, for example a standardized effect corresponding to a partial correlation or Cohen's f^2 , with a specified α level. In practical terms, this often implies a fairly large N . Our current rough estimates suggest something like $N = 60$ to 100 sessions in a within-subject design, where each person provides many trials, or $N = 120$ or more participants in a between-subject design, may be needed to reliably detect an effect of the size we expect under rigorous cross-validation. For neural studies, where measurement noise is higher, we might need the higher end of that range or more, especially if looking for effects like $\Phi \times H$ interactions. Note that overfitting can make small datasets look strong. The power goal is not exploratory signal fishing but high-confidence detection. QS is a high-level hypothesis, so initial critical tests should be well powered and confirmatory.

Conducting these power and calibration steps ensures we do not fall into the trap of underpowered study designs yielding false negatives, meaning no QS is found because we did not have enough data, or false positives, meaning we find something in a small sample that later does not replicate. We aim for the first critical tests of QS to be definitive if possible.

5.5.10 Reporting Standards (What to Publish)

When it comes time to publish results, especially about something as potentially foundational as a QS-residual, transparency and completeness are essential. We commit to the following reporting standards.

- Full specification of models and preprocessing: We will describe the nuisance backbone in detail, including every variable included and why. If certain variables were left out, we explain why, for example because they were measured but found redundant or unreliable. We will also share the exact code or equations used, such as how Φ was computed from raw data. The same goes for priors in Bayesian models or any regularization. Preprocessing steps such as how we handled outliers, how we normalized physiological signals, and how we documented trial-level exclusions must be fully specified. Ideally, provide a supplement or repository with analysis scripts so others can re-check.
- All model variants and comparisons: Present not only the final model but intermediate models. For example, Model A backbone only, Model B backbone plus Φ , Model C backbone plus Φ plus horizon term, Model D backbone plus Φ plus horizon term plus social penalty. We report predictive performance metrics, such as LOO, AIC, or WAIC, for each model and where appropriate a statistical comparison such as likelihood ratio tests or Bayes factors between nested models. This shows the incremental value of each component. If adding a QS term does not improve the model, that is reported, not swept aside.
- Effect size estimates with uncertainty: For each QS-related coefficient, for example γ_{Φ} in behavior or γ_{Φ} in neural, report the estimate and a confidence or credible interval. Also report how sensitive results are to analysis choices, for example whether results change when an alternative set of nuisance terms is used, whether the effect depends on a dichotomous versus continuous horizon variable, or whether the effect vanishes on certain subsets of trials. Include negative control results, which should show null effects where expected, and report any cases where the QS effect went away, which might indicate a boundary condition. Present a nuanced picture rather than a single binary significant or non-significant label.

- Replication attempts: If we have data from multiple labs or run a follow-up, include those results even if the replication was weaker. The goal is an honest accounting of consistency. Also, if alternative analyses such as using a different modeling approach were performed by us or an independent group, report those as well. We will also share summary statistics that allow independent meta-analysis or re-checking of core claims, for example per-subject QS effect sizes or anonymized trial-level data with sensitive fields removed.

By adhering to these standards, we make it easier for the scientific community to critique, replicate, and build on our findings. Given the extraordinary claim that a law of fairness might be operating in the brain, the evidence must be extraordinarily open and robust.

5.5.11 Typical Failure Patterns and How to Interpret Them (*Fail-Pattern Box*)

Even with careful methods, experiments do not always support the hypothesis. Here we list common failure patterns we might encounter when hunting for QS-residuals, and what each would imply.

- QS effects vanish after adding a previously omitted covariate: Suppose our initial analysis found a significant Φ effect, but then we realize we forgot to include a subtle but important nuisance variable, perhaps a socioeconomic status effect or a specific task variable. Once we add that, the QS effect drops to zero. Interpretation: the QS effect was likely an artifact of a missing piece in the backbone. The remedy is to revise the backbone model and be cautious about over-claiming. This does not kill QS entirely, but it means that dataset did not demonstrate anything beyond a standard effect.
- Φ is significant but the $\Phi \times g(H)$ interaction is null: We might find that Φ influences choices in general, people lean toward more compensatory moves, but this effect does not get stronger as horizon shortens. Interpretation: this outcome suggests what we are seeing is more of a general tendency toward balance rather than a strict horizon-driven constraint. People might generally prefer reparative acts and avoid irreparable harm, which is interesting, but it could be explained by ordinary foresight or virtue, not a law that intensifies in the endgame. It would weaken the argument that LoF operates as a time constraint. It would make QS look more like a background preference that does not necessarily guarantee closure by the end.
- Social penalty λ shows no effect despite real congestion: If we run a multi-agent or resource competition experiment where objective congestion is present, but the λ term has no effect, then the social coupling part of QS is not manifest. Interpretation: either our specification of the social effect is wrong, perhaps we

modeled it poorly or assumed linearity when it is nonlinear, or in that domain, people do not coordinate via the mechanism we are proposing. Cultural or communication factors may override QS, or individuals may act selfishly in ways QS does not correct. This null result would require us to either refine the social component or acknowledge a boundary condition. It does not falsify QS overall, but it suggests the social coordination piece is weaker than claimed.

- No evidence of dream mediation: If controlling for dream affect D does not reduce the correlation between day's ledger change and next-day behavior, then dreams might not be carrying the load we expected. Interpretation: the counterweight function could be occurring through waking processes rather than dreams. For instance, people may solve their imbalances through daily social interactions or reflection, leaving little for dreams to do, or our measures of dream affect may have been too crude. This pattern would prompt us to revisit assumptions and experimental designs. It would mean that in our sample, dreams were not effectively engaged as counterweights, which could be a real challenge for QS if it generalizes, or it could indicate we need better dream measurement.

Each of these failure patterns, if observed, does not necessarily kill the whole theory instantly, but each is a strike. We would interpret them, explore fixes or alternative explanations, and delineate limits. They help specify when and where QS may operate, and they guide the next iteration of hypotheses.

5.5.12 Minimal Milestone for “QS-Residual Observed”

What would success look like? Before crowning QS as “observed,” we set a high bar. At minimum, a study should demonstrate all of the following to claim credible detection of a QS-residual:

1. Neural signatures: We need to see the predicted QS pattern in the brain. This means, for example, in value-sensitive regions such as vmPFC or OFC a significant positive $\beta\Phi$ indicating greater activity for options that improve future balance, and concurrently in inhibitory or control regions such as rIFG or subthalamic nucleus evidence of braking for low- Φ options, such as increased stopping-related activity when someone contemplates a highly non-compensable action. The ACC should show sensitivity to horizon-scaling terms, for example increased conflict signals for short-horizon high-stakes choices, and to shared-resource load terms when options impose high λ -weighted resource costs. Importantly, these signals should each explain variance beyond standard nuisance factors defined in the backbone. For instance, adding Φ and $\Phi \times g(H)$ regressors should significantly improve cross-validated model fit for those ROIs.

If we observe only vmPFC sensitivity to Φ without corresponding control or horizon-sensitive signals, that constitutes partial support but not the full predicted circuit signature.

2. Behavioral effects on choices and RTs: The QS-augmented model should predict choices and reaction times better than the backbone model. This could manifest as improved prediction accuracy or reduced log-loss when horizon interactions are included. Specifically, we expect conditions in which horizon shortens to show increasing weight on Φ . For example, someone might initially start down a risky path when time seems ample but shift away from it as the end of the task nears if Φ is low. Likewise, conflict between immediate reward and future consequence should resolve more strongly toward compensable options under short horizons. Quantitatively, including Φ and $\Phi \times g(H)$ should reduce prediction error or increase pseudo- R^2 in logistic models beyond what the backbone explains, particularly in late-game or short-horizon trials.
3. Robustness and replication: The above findings must survive reasonable checks. That includes alternative nuisance specifications, different cross-validation splits, and subsample analyses. Independent replication in a second dataset or laboratory is strongly preferred. If the effect appears only under one specific analytic choice or in one sample, that is insufficient. By the milestone of “QS-residual observed,” we mean a strong, repeatable divergence from conventional models that survives cross-validation and specification stress tests.

Meeting all three criteria would allow us to cautiously say: yes, there is residual structure in brain and behavior consistent with a constraint favoring compensable paths. At that point, we argue QS moves from hypothesis toward empirically supported mechanism, even if interpretation remains open. Anything less, such as neural-only or behavioral-only evidence, or effects that collapse under alternative analyses, must be marked as tentative and categorized under “tendency” explanations rather than proof of a law-level mechanism.

5.5.13 Where we go next:

A mechanism that cannot lose is not a mechanism. The next section lists concrete failure cases that would count against QS, including no narrowing after large ledger shocks, absence of a residual control signature, or patterns incompatible with a feasibility constraint.

5.6 What Would Falsify QS?

A constraint hypothesis is only scientific if we can state clearly what would prove it wrong. This section enumerates decisive, preregisterable tests that would falsify the Queue System (QS) as the implementation of the Law of Fairness. We specify measurable outcomes, directions of effect, and analysis plans, so that if any of these outcomes occur under adequate power and proper controls, they would force us to either retreat to a weaker tendency interpretation or abandon the QS idea entirely. In short, we describe how QS could fail, the failure modes that, if observed, show that nature is not following the Law of Fairness via QS.

5.6.1 Strong Falsifiers (Single-Study Decisive Outcomes)

Each item below is a critical experiment or observation that could on its own deal a serious blow to the QS theory. Accompanying each is how we would test it and what it would mean.

1. No QS-residual in the brain after exhaustive controls:

- *Test:* Using neuroimaging (fMRI, EEG, etc.), we run the full model including all nuisance factors (utility, conflict, risk, habit, learning signals, fatigue, etc.) and then add the Φ -family terms (Φ , $\Phi \times g(H)$, social penalty). We examine hypothesized QS-related regions such as vmPFC, ACC, rIFG, insula, and associated circuitry such as subcortical nodes. If none of these regions show any improvement in predictive power or systematic activity associated with QS terms, for example adding Φ and $\Phi \times g(H)$ provides no out-of-sample gain in model fit and all QS coefficients' posterior intervals include zero, then QS has no neural trace.
- *Interpretation:* This would mean QS lacks unique explanatory variance in the brain. After accounting for known decision processes, there is no detectable residual pattern consistent with a fairness constraint. That strongly suggests QS, as a distinct process, does not exist at least in any straightforward measurable way.

2. No horizon scaling where it ought to appear:

- *Test:* Design an experiment where time horizon H is manipulated while holding immediate utilities constant. For example, vary how much time remains, such as telling participants “this is your last chance” versus “you will have more chances later,” while keeping short-term payoffs equal. If QS is correct, short-horizon conditions should increase the weight on Φ

relative to long-horizon conditions. Concretely, the $\Phi \times g(H)$ interaction should be significant across measures such as choices, reaction times, and neural responses. If there is no measurable difference at all in behavior or brain activity between short and long horizon contexts, or no intensification near the end, that contradicts the horizon mechanism.

- *Interpretation:* If short-horizon individuals behave identically to long-horizon ones in compensability contexts, then the endgame intensification idea is unsupported. Observed adaptation may reflect general time effects rather than a constraint that strengthens as H shrinks. The Law of Fairness would then resemble a static preference rather than a dynamic constraint.

3. No counterweights during sleep:

- *Test:* We examine the link between daily emotional imbalance and dream content, as well as the effect of REM sleep on next-day adjustment. If our Day–Night inversion study (Section 5.4.5A) yields nothing – i.e., the correlation between daytime ledger change ΔL_{day} and dream affect D is around zero (people’s dream emotional tone has no reliable relation to their day’s experiences), and moreover adding a “time pressure” prime doesn’t change dreams – that would refute the idea that dreams systematically compensate. Additionally, consider the REM manipulation (Section 5.4.5C): if suppressing REM sleep or blocking dream recall has no effect on next-day choices or mood (for instance, people deprived of REM are *just as good* at bouncing back or seeking reparative actions as those with normal REM), it means the supposed work of dreams can be skipped with no consequence.
- *Interpretation:* This would mean dreams are not functioning as low-cost counterweights. It could be that dreams are mostly epiphenomenal or serving other functions (like memory consolidation) unrelated to QS. If true, QS loses one of its most novel predictions (that even when awake we’re “offline” balancing), and we’d have to accept that any balancing is done through waking life only. The Law of Fairness would then have to operate without the night shift – or not at all if waking evidence is also weak.

4. No social coupling under resource congestion:

- *Test:* Take scenarios of resource contention, either in controlled experiments or observational studies. For example, in a multiplayer game

where only a subset can get a reward (introducing a queue or competition), or in hospital data where many patients vie for limited ICU beds. QS predicts coordinated behavior: individuals' choices or outcomes should show co-movement with the congestion metric λ , and brain regions like ACC and rIFG should encode the summed penalty $\sum \lambda \Delta r$. A falsification occurs if we find nothing of that sort: no statistically significant synchrony in behavior (everyone just does their own thing independent of each other), no ACC activation relating to social load, and no sign that short-horizon individuals get any priority when competition is high. In short, when we measure $\lambda_r(t)$ (the resource load over time) and each person's behavior, there's zero correlation or pattern linking them.

- *Interpretation:* If admissible options don't actually depend on others' usage of shared channels, then QS misses a "core real-world feature" it claimed. The theory of LoF heavily implies an interdependence – we affect each other's ability to compensate. No evidence for that means maybe fairness in one life doesn't really couple to fairness in another via the mechanisms proposed. Perhaps balancing is a purely individual affair, or any group-level coordination is due to explicit social norms rather than an intrinsic constraint. It would force us to drop or radically alter the social component of QS.

5. Admissible-set leakage near closure:

- *Test:* Look for cases where the guardrails utterly fail in the precise situation they're supposed to be strongest: near the end of a life or other final horizon. For instance, track individuals (in detailed case studies or datasets) who have very little time remaining *and still* have possible ways to make amends or get relief. If we find multiple instances where such individuals choose blatantly low- Φ trajectories – essentially catastrophic decisions – *without any evidence of internal struggle or braking*, that's a serious violation. Concretely, imagine observing terminal patients who, despite having the opportunity for closure or comfort, engage in self-destructive behavior (substance abuse, alienating loved ones, refusing any help) and we detect no psychological resistance, no last-minute change of heart, no stall – just a straight dive into imbalance. If this happens repeatedly in well-observed contexts (with no extenuating factors like sudden mental deterioration), it indicates QS is not in effect. In lab terms, an analog would be an experiment where participants nearing the end of a

task can either salvage some points or make a risky all-or-nothing bet that if lost leaves them irrecoverably in the red – and many go for the bet with zero hesitation and end in ruin, contrary to QS’s prediction that near the end they’d avoid irrecoverable harm.

- *Interpretation:* This is basically catching QS asleep at the wheel. If people can drive their life (or game) off a cliff with no internal alarm kicking in at the final moments, then the supposed constraint fails “when it matters most.” One could argue perhaps those instances are due to other pathology (e.g., severe psychiatric conditions) – and indeed we’d have to scrutinize that. But if even mentally competent individuals do this, it’s a lethal blow to the idea of a robust built-in fairness guardrail. It would suggest either the Law is false or it’s so weak as to allow major violations, which undermines calling it a Law.

6. A rival model reproduces all signatures with fewer assumptions:

- *Test:* Develop or take an existing model that has no fairness constraint in it at all – for example, a combination of predictive coding (minimizing surprise), risk aversion, and fatigue or homeostatic drive + learning – and fit it to the same data. If this rival model can match the performance of the QS-augmented model on neural and behavioral fits, and can also explain qualitatively the phenomena of interest (day-night mood changes, social coordination, etc.) *without referencing Φ , horizon, or λ* , then QS is unnecessary. We essentially ask: is there a simpler or more parsimonious explanation for everything we’ve ascribed to QS? This could be tested by formal model comparison: if a constraint-free model has equal or better goodness-of-fit (within cross-validation) and comparable ability to predict new observations, then adding QS doesn’t buy us anything. Or conceptually: maybe a “risk management + resource rationality” theory predicts that near end-of-life people change behavior due to shifting risk tolerance and opportunity cost of time, not fairness – if that covers all evidence, QS isn’t needed.
- *Interpretation:* By Occam’s Razor or the best-system principle, if all observed signatures (like horizon effects, sleep inversions, co-movement) can be explained by stitching together known mechanisms (each perhaps domain-specific) rather than invoking a new global fairness mandate, then we should prefer the simpler explanation. QS would then be an overfitting—an elaborate story we don’t need. We would drop it from our

framework and say, “Nature might not have a fairness law; it could just be a mix of ordinary processes.”

Any one of these outcomes, if convincingly demonstrated, would be enough to call QS into question. They each target a central pillar of the theory. We have, in fact, incorporated many of these as pre-specified “fail conditions” in our research plans (see Section 5.1.6), precisely to avoid clinging to the theory if the data don’t support it.

5.6.2 Decisive Experiment Set (Minimal Program)

We can outline a minimal set of experiments that, collectively, would test the major predictions of QS. Each experiment (E1–E5) comes with a specific criterion for falsification. If QS is correct, all these experiments should yield results consistent with the theory; if even a couple of them fail decisively, the theory is in trouble.

- E1: Horizon-manipulated “Repair vs. Indulgence” Task (fMRI + Behavior): Participants make choices between options that either help “repair” their ledger (e.g., doing a responsible or kind action that might not be immediately rewarding) or “indulge” at the cost of future consequences (e.g., a pleasurable action that could cause harm later). We experimentally manipulate the *time horizon* within the task: sometimes people believe it’s the last round or that they have very few chances left (short horizon), other times they believe many rounds remain (long horizon). We preregister that we’ll include QS regressors (Φ for each option, an indicator or inverse for short horizon, and $\sum \Delta r$ if there’s a shared resource aspect) in both neural and behavioral models. QS expects: in short-horizon conditions, the brain should show stronger signals associated with compensability – specifically: vmPFC value signals should weight Φ highly (so high- Φ options get a big neural value boost) meaning $\gamma_\Phi > 0$; rIFG or STN (inhibitory regions) should activate more for choosing low- Φ options under short horizons (basically evidence of a brake, which means a $\Phi \times H^{-1}$ interaction where those regions fire more when horizon is short and Φ is low, i.e. effectively Φ has a negative effect on these regions that is amplified by short H); ACC should show increased conflict or control signal correlated with the social penalty term ($\sum \Delta r$) if the task involves resource sharing. All these neural effects should correlate with behavior: the model with Φ and horizon should predict choices better (LOO/WAIC improvement). *Falsifier:* If we find that any two of the three critical brain region predictions fail – for instance, vmPFC shows no Φ effect and rIFG shows no braking – even with sufficient power and proper controls, then QS’s core neural mechanism is not supported in this context. Failing two out of three would mean it’s not just a fluke in one region; it suggests a broader absence of QS dynamics.

- E2: Congestion Paradigm (Behavioral + EEG/fMRI): Set up a scenario where individuals make choices in the presence or absence of resource contention. For example, imagine an online platform where people can request help from a common pool (like asking questions to a limited number of experts). In high congestion condition, many requests are being made (λ is high); in low congestion, few requests (λ low). We also include a manipulation of whether participants know others' needs (to engage the social aspect). QS predicts: when resource contention is high, we should observe menu co-movement – people collectively avoid heavy-draw options and switch to alternatives, perhaps measurable as an increase in some index of synchrony or as a significant effect of $\sum \Delta r$ on choices (everyone's choices shift when λ changes). Also, short-horizon individuals (if some participants are in a “urgent” condition) should still take the heavy resource (horizon-priority), whereas long-horizon individuals back off – an interaction to test. In the brain/EEG, we might see synchronized changes in frontal theta or ACC activity correlating with group load. *Falsifier:* If we see no evidence of coordinated shifting (no co-movement) and no neural or behavioral effect of the social penalty regressor, then the QS social coordination claim fails. Essentially, if people don't respond to congestion in the predicted way – each just pursuing their utility regardless – then QS's shared resource mechanism is not operating.
- E3: Day–Night Counterweight Study (Sleep Lab + Dream Reports + Next-Day Choices): This combines elements from Section 5.4.5. We recruit participants to undergo the day–dream–next-day cycle in a controlled way. Day 1: induce some emotional drift (maybe a challenge or stress to create ΔL_{day}). Night 1: record sleep (polysomnography) and collect dream reports. Morning Day 2: measure their choices in some tasks that allow compensatory actions. We also include a horizon prime for some (e.g., tell half of them some subtle “time is short” message before sleep, perhaps to see if that increases α). Optionally, do a REM suppression for half the night and let it rebound in second half (within-subject). QS expects: a negative slope between ΔL_{day} and D (bad day -> negative dreams) and that slope gets steeper if a horizon prime is present (or for those high in trait urgency). Also, if REM is experimentally suppressed, that slope should weaken (as they can't dream properly). Next day, those who had strong counterweight dreams should show a boost in high- Φ choices (like more willingness to do reparative acts) compared to those who didn't. If REM was suppressed, that next-day boost should be smaller; if REM rebounded, maybe an overshoot of compensation. *Falsifier:* If across participants we find the dream affect vs. day's stress slope is essentially zero (or positive, which would be opposite), *and* horizon cues make no difference, *and* REM manipulation changes nothing about next-day behavior, then

the whole counterweight notion is unsupported. We'd conclude dreams are not playing the QS role.

- E4: Sequential Persistence / Stall Test (Multi-Step Risky Policies): In a lab task, present participants with scenarios where they can take a sequence of actions (e.g., keep investing in a gamble or escalate a conflict) and at any point they can stop. Some sequences if continued to the end lead to a disaster (large negative outcome) that cannot be recovered, so continuing them too far is low-Φ. Other sequences might be safe or reversible. We vary how many steps remain (the horizon within the task). QS predicts: Low-Φ action sequences will be more likely to “stall out” as the horizon gets shorter. For example, if someone only has a couple of moves left and they are on a dangerous path, they should be more likely to quit or switch to a safer action compared to someone who had a lot of moves left who might continue that path. Essentially, a strong interaction: shorter horizon -> higher probability of terminating a low-Φ sequence. We measure things like the proportion of sequences where the participant voluntarily stops vs. goes all the way.

Falsifier: If we find no difference in behavior with horizon – say participants continue risky sequences at the same rate even when time is almost up – so there's no selective stalling for low-Φ sequences, then QS's horizon effect in sequential decisions is not there. In other words, the “stop taking detours when almost out of time” intuition would be falsified.

- E5: Naturalistic Telemetry (Weeks-Long Real-World Monitoring): Track individuals over an extended period (using wearables, smartphone data, diaries). We look for broad QS patterns in life: As people near a significant ending (could be end of observation period, or an event like graduation, moving, etc.), do we see their “admissible set” of activities narrow and tilt towards compensatory ones? We'd predefine indices like: *repertoire breadth* (variety of different activities or places they go each day – predicted to shrink if they focus on essentials), *stickiness asymmetry* (whether negative moods or behaviors dissipate faster or slower than positive ones – QS might predict a tendency to not let negative spirals continue unchecked as time goes on), *menu co-movement* with shared resource metrics (if we have groups, like a family, do their patterns sync up when one member's needs spike?), and *priority for those with short horizons* (if some cohort members become constrained, do others yield resources to them). We gather things like communication logs, health data, social interactions to gauge these.
Falsifier: If none of these expected patterns show up despite high-quality measurement, that's a blow. For instance, if as people approach an ending there's

no reduction in the variance of their accumulated affect (we'd expect ledgers to perhaps converge or at least not go wildly apart), if there's no difference in how long negative states last vs. positive (we might expect some asymmetry like interventions curtail negative episodes), and no sign of synchronized adjustments among connected individuals, then QS isn't showing up in real life. Essentially, if people's life trajectories remain just as idiosyncratic and uncorrected at the end as at the beginning, then the law-like regularity isn't there.

This set of experiments covers individual decision-making, social coordination, sleep, sequential behavior, and longitudinal observation – the main arenas where QS would manifest. Failure of two or more of these (especially E1, E2, E3) in replication would be enough for us to seriously doubt QS as the correct model (see next section on decision rules for retreat).

5.6.3 Boundary Fail Conditions (Where QS Could Be “True but Trivial”)

Not every failure means “QS is false”; some could mean “QS is technically there but negligible.” We outline boundary conditions that would effectively make QS a trivial curiosity rather than a meaningful law:

- Microscopic effect sizes: Suppose we do detect QS-consistent signals, but they are extremely tiny – e.g., in fMRI, the vmPFC activation to Φ is only 0.1% above baseline (virtually indistinguishable from noise), or in behavior the presence of a short horizon changes choice probabilities by only 0.5%. If after thorough study we see that QS effects exist but are an order of magnitude smaller than other factors (like much smaller than typical utility effects) such that they never actually change decisions in practice, then QS might be “true” in a technical sense but trivial in impact. We'd essentially have a very weak constraint that doesn't guarantee anything – lives could still go wildly unfair because the supposed law barely nudges choices.
- Population-specific or fragile effects: If QS signatures only appear in very limited circumstances – say only in young adult samples but not in older adults, or only in lab tasks but not in more realistic settings, or only when people are somewhat stressed but not when very calm. For instance, maybe you find horizon effects in a student sample cramming for exams, but zero effect in working professionals planning retirement. Or dream counterweights might show up in people who journal about dreams but not in others. If any observed QS effects vanish under slight task changes or in different subgroups, that means it's not a universal law, just a peculiarity of certain conditions. A law-level claim requires generality; if it's that fickle, it's not really law-like.

- Analysis or meter dependence: This is about reproducibility and measure choice. Suppose one particular way of analyzing data yields a QS effect (e.g., using one type of questionnaire or one specific fMRI analysis pipeline), but if you use an equally reasonable alternative, the effect disappears. For example, QS appears if you compute “compensatory choices” in one arbitrary way, but not if you use a slightly different definition; or an EEG measure shows something but an fMRI doesn’t, or vice versa, and there’s no good reason for the discrepancy. If the evidence for QS is highly dependent on the exact analytical method or instrument, it might suggest we’re picking up noise or an artifact. True laws should be detectable by multiple methods.

If these boundary conditions hold up in replication – meaning multiple attempts and analyses consistently find only minuscule or highly fragile effects – then even if we cannot say QS is completely false, we’d have to conclude it’s not a “law-level” mechanism. It would be more appropriate to classify it as a marginal phenomenon, not something that reliably governs behavior. Essentially, “the juice wouldn’t be worth the squeeze” – QS might exist but matter so little that it doesn’t meaningfully constrain life outcomes.

5.6.4 Rival Frameworks That Could Win

We must acknowledge that there are alternative theoretical frameworks which could account for much of the data we attribute to QS, without invoking any cosmic fairness principle. If one of these frameworks ends up explaining the evidence better, QS will have to step aside. Some contenders:

- Homeostasis + Opponent Processes + Standard REM Functions: Perhaps the simplest null hypothesis is that people have a psychological homeostasis: after extreme experiences, internal opponent-process mechanisms kick in to bring affect back to baseline (Solomon and Corbit’s theory, for example). Couple that with known REM sleep functions like memory consolidation and fear extinction (which aren’t about fairness per se, just general emotional regulation). This combo might explain why people often bounce back after hardships (it’s just the opponent after-effect) and why dreams sometimes mirror waking concerns (standard memory processing). End-of-life shifts like seeking closure might be explained by basic social and emotional needs becoming salient as one prepares for death (nothing to do with a ledger, just priorities shifting naturally). If careful experiments show that all QS-like observations map onto what these familiar processes predict, then adding a fairness constraint is unnecessary.

- Free-Energy (Predictive Coding) + Risk/Uncertainty + Arousal Modulation: Predictive coding theory says the brain tries to minimize surprise or prediction error. One could conceive that many behaviors we think of as “making things fair” are actually just about reducing psychological uncertainty or distress. For instance, a person might resolve unfinished business at end-of-life not due to fairness, but to reduce unpredictability or anxiety about interpersonal relationships (i.e., minimize expected surprise in social interactions). If you augment a predictive coding model with standard risk aversion (don’t take options that have big unknowns or potential big downside) and arousal regulation (people avoid sustained high stress), you might mimic QS. For example: short horizon could raise arousal (thinking about death), which makes certain risky indulgences feel extra aversive (since they’d spike arousal more), leading to exactly the behavior QS predicts, but purely from a surprise-minimization standpoint. Similarly, dreams might be reframed as the brain’s way of reducing prediction errors about emotional events, not about compensation. If such a model, when fleshed out, fits data as well as QS does, it’s a strong alternative.
- Resource-Rational Reinforcement Learning (RL) with Queueing Costs: In AI and cognitive science, resource-rational models suggest agents account for computational or resource costs in their decision-making. Imagine a reinforcement learning agent that, on top of reward, has a penalty for using too much of a scarce resource or time. That agent would naturally “queue” its actions: if many others want the same resource, it might take an alternative to avoid waiting (not because of fairness, but to avoid time cost). It might also behave differently when time is short because the opportunity cost of time changes. Essentially, if you add a general cost for waiting and a cost for risk, a sufficiently complex RL model might show horizon-dependent choices and avoidance of congested resources – basically doing what QS does but attributing it to rational optimization of resources. This would frame those behaviors not as moral balancing but as efficient use of time and resources.

If any of these (or other) rival frameworks can quantitatively match the entire signature set that QS aims to explain – horizon effects, social coupling, sleep counterweights, stall patterns, final ledger compression, etc. – and do so with equal or better predictive accuracy using fewer free parameters or assumptions, then by scientific standards we should favor the rival. QS would then be viewed as an over-specified story; the simpler explanation would take precedence.

We remain open to this outcome. In fact, our research plan includes running such rival models in parallel (see the note on adversarial collaborations below and also

Section 5.1.6 fail conditions: “rival accounts match all signatures”). If a rival wins, it’s progress for science: we’d have learned how those phenomena come about without needing a new law.

5.6.5 Partial-Support Patterns (What Helps but Does Not Fully Save QS)

Now, what if we get some encouraging results but others not so much? Here are patterns that would be *consistent with QS in part*, yet still insufficient to declare the law intact:

- Isolated vmPFC Φ effects without rIFG/ACC modulation: Suppose our experiments show that the brain’s value center (vmPFC) reliably values high- Φ options more (we get that positive γ_Φ), suggesting people do neurally recognize compensable vs. non-compensable differences. But we do not see the complementary control signals – the rIFG doesn’t selectively inhibit low- Φ impulses, the ACC doesn’t show extra conflict for non-compensable choices. This would mean the “gas pedal” of QS is there but the “brake” is not. It’s like people recognize the good path but don’t necessarily stop the bad path. That’s helpful (it indicates some fairness weighting in the mind) but not enough to enforce the law, since without brakes one could still go off track.
- Dream counterweights present, but no horizon dependence: We might find that indeed bad days lead to bad dreams and good days to good dreams (supporting a basic inversion), confirming part of QS’s sleep story. However, if that effect is the same regardless of whether time is running out or not – in other words, $\alpha(H)$ in the equation from Section 5.4.2 is basically constant, not increasing as H shrinks – then QS’s specific claim about urgency is unsupported. Dreams might be balancing things somewhat, but not in a way that “knows” about the end of life. That scenario would suggest maybe a general emotional homeostasis in dreams rather than a targeted fairness mechanism.
- Social co-movement observed, but no horizon-priority or neural social coding: It could be that in group experiments we do see people adjusting together when resources are scarce (so co-movement happens, indicating some coupling). But if we don’t see the *priority* aspect – e.g., everyone just equally cuts down usage, rather than short-horizon folks being allowed more – then QS’s claim of how that coordination works is incomplete. Additionally, maybe we see behavioral coordination but when we look in the brain, we don’t find ACC or insula encoding the social penalty; the coordination might be happening via explicit communication or instructions rather than an intrinsic QS process. So, it’s partially confirming (there is coordination) but not via the QS-proposed mechanism (no sign of λ -based priority).

- Better in-sample fits, but no out-of-sample gains: It's possible we find that adding QS terms can explain the data we collected really well (the model fits look great, p-values small), but when we try to predict new data or do a cross-validation, the advantage vanishes. This suggests that QS might be overfitting to noise in the initial data – capturing idiosyncratic patterns that don't generalize. It's a classic scenario where a model appears to work until you test it on new examples. If QS falls into that trap, then even if it “worked” on some dataset, we wouldn't count it as validated; it would still be provisional.

All these partial supports keep QS in the conversation – they show it's “interesting”, maybe even on the right track – but they demote it to a hypothesis or tendency rather than a firm law. We'd say something like: “There might be a fairness-related factor in decision-making, but it doesn't consistently do everything the Law of Fairness would require.” QS might join the ranks of many psychological constructs that are suggestive but not ironclad.

5.6.6 Decision Rule for Retreat

We believe in precommitting to criteria for giving up on a cherished hypothesis. Here is our explicit decision rule for QS: If two or more of the strong falsifiers from Section 5.6.1 are confirmed (replicated) in well-powered studies, we will withdraw the claim that QS provides a guaranteed fairness constraint in life. In that case, we would likely reformulate the Law of Fairness as perhaps a heuristic or tendency – something that often, but not always, influences behavior (if there's still some effect left).

If three or more strong falsifiers replicate, that's effectively game over for QS; we would abandon QS as an explanatory construct for the Law of Fairness. We'd then conclude that whatever balancing behavior is observed can be better explained by other mechanisms or is not reliably present at all.

For example, let's say experiments show no neural QS-residual, no horizon effect, and dreams have no impact (three falsifiers) – that would definitively refute the machinery we proposed. Or maybe neural QS is found but dreams and social coupling both fail and a rival model covers the rest – that mix could also hit the “two or more” rule depending on which ones. The point is we won't cherry-pick: multiple major failures means time to concede.

By specifying this ahead of time, we avoid the temptation to shift goalposts or rationalize away disconfirmations. It's an all-too-common problem that theorists, having invested in an idea, keep it alive with ad hoc fixes. We aim to do the opposite: if nature clearly tells us “no, it doesn't work like that,” we'll listen and pivot our theory accordingly.

5.6.7 Data Quality, Ethics, and Adversarial Checks

Finally, a note on how we'll conduct and scrutinize this research, because extraordinary claims require not only extraordinary evidence but also exemplary ethics and rigor:

- Sufficient power and preregistration – non-negotiable: We will not rely on underpowered studies or post-hoc storytelling. Every critical experiment mentioned (E1–E5, etc.) will be preregistered with planned sample size justified by power analysis, analysis pipelines set in advance, and clear criteria for success/failure. We'll use robust validation like blocked cross-validation and holdouts, as discussed. We will also include negative controls (like sham congestion, placebo horizons per Section 5.5.7) by design in these studies so that a reader can be confident any positive result isn't a fluke. Essentially, we aim to leave “no wiggle room”: the design either yields the effect or it doesn't, and if it doesn't, that's that.
- Ethics: comfort and dignity override data collection. Testing QS could involve vulnerable scenarios (end-of-life, psychological stress, etc.), so we adhere strictly to ethics. We do no trauma induction, period. If we need to study severe pain or loss, we will use clinical cohorts by consent (people who unfortunately experience those anyway, such as patients, but only with full ethical oversight and if they willingly participate). Animal studies, if any, will follow the 3Rs (and frankly QS is a very human psychology concept, so animal work might be limited to basic neural circuit questions). Importantly, if at any point a participant's well-being is at odds with our data needs, well-being wins. To reiterate our guiding principle: Relief is a systems variable; comfort and dignity override data collection. In practical terms, we would never, for example, withhold analgesics or counseling from someone just to see how their “ledger” plays out untreated – that would be grossly unethical and scientifically invalid too. End-of-life observations, if any, will be purely observational or integrated with care (and only with consent) – never would we encourage someone to forgo comfort care for the sake of measurement. The ethos is: *the hypothesis must never trump humanity*. If QS is real, it should show up in ethical observation; if it doesn't in that context, too bad for QS.
- Adversarial collaborations and blind analysis: We plan to involve skeptics in the analysis process. For example, we might invite a proponent of a rival theory (like a noted predictive coding researcher) to analyze the data their way. We can do “code swaps” where our team and an opposing team both analyze the same data independently and compare notes. We'll also consider preregistered *blinded analyses* – where the data is provided without condition labels (randomly shuffled labels that are revealed only after analysis choices are made), to avoid bias. All

results, especially failures to find QS effects, will be reported as prominently as successes. If we find nothing, that itself is valuable information given the scope of the claim. We won't bury non-significant results; in fact, those are the ones that directly test the theory's falsifiability. Our publications or reports will highlight, for instance, "Experiment X did not find the predicted horizon effect, despite Y power," etc.

5.6.8 Where we go next:

If QS trims options to keep compensation possible, shrinking time should magnify that pressure. Chapter 6 takes up horizons directly, introducing H_t and the shadow price λ_t as a way to formalize why end-game balancing should intensify if the constraint is real. We will only consider QS validated if it produces clear, distinctive signatures – *horizon-sensitive, socially coupled, sleep-assisted patterns* – and those survive the gauntlet of rigorous controls and out-of-sample tests. Conversely, if those signatures are consistently absent when they should be present, it's not a minor blemish we can brush aside; it is a falsification of the QS hypothesis. We have laid out exactly what those crucial signatures and tests are, and we stand prepared to let the evidence speak.

Chapter 6 — Time Horizons and the Shadow Price

Think about how differently we act when we’re running out of time. A student in the last week before final exams zeroes in on studying, ignoring parties or new hobbies. A patient who’s told they have only a few months left to live suddenly feels an urgency to reconcile with estranged family, to say what needs saying. Deadlines, countdowns, and finite horizons have a way of sharpening priorities. Why is that? This chapter explores the idea that as the time to make things right grows short, the internal cost of each choice changes dramatically. If the Law of Fairness is in play, the final stretch of any life or any significant period of life is like the endgame in chess: mistakes are costlier, and the moves that lead to closure become immensely valuable. We introduce the notion of a shadow price on time remaining — an implicit weight or cost that increases as the horizon decreases. Throughout this chapter, “horizon” refers to effective remaining compensable time H_t as defined in Section 6.4, not merely chronological clock time; all claims concern changes in decision weighting conditional on measured or credibly manipulated H_t . In simple terms: the less time you have, the less wiggle room you have for detours, so the system responds by weighting options more heavily toward those that promote balance and closure.

At a high level, the logic is straightforward. Imagine two people each tempted to send a bitter, hurtful message in an argument. One person is young and healthy with years ahead to mend any hurt feelings. The other is boarding a plane to see a dying parent one last time — this might be their final opportunity to speak to that parent. For the first person, indulging in a nasty remark is possible — they might later apologize and live on. For the second person, that same cruel message is almost unthinkable; it does not even come to mind with the same force. It’s as if an internal circuit says, “No, not now — you may never get the chance to fix this.” The option slides out of consideration, replaced by gentler alternatives, perhaps an apology or a loving message, that feel right for the moment. Nothing mystical changed in the second person’s character; what changed was the time horizon. With virtually no future in which to make amends, the internal cost of a wrong move spikes so high that the hurtful action effectively drops out of the menu. In Queue System terms: the compensability Φ of a hurtful action went deeply negative and the horizon weight $\beta(H)$ magnified that negativity, showing the action off the mental queue. In earlier notation, this is the $\Phi \times g(H)$ horizon-scaling term, with $g(H)$ often instantiated as $1/(H+\delta)$ so the scaling remains finite as H approaches 0. A conciliatory action, by contrast, had a high Φ and rose to the top of the queue.

Now layer in a shared-resource scenario: suppose a small hospital has one ICU bed open and several patients who need it. Everyone’s options narrow somewhat — they all feel a

nudge to consider alternatives like less intensive treatments or waiting for a spot, an effect of a rising resource cost $\lambda_r(t)$ on that ICU bed. But for a patient who truly cannot wait because their horizon is extremely short, the system quickly makes an exception: the high-need patient does not feel the same nudge to step aside. In fact, they may feel a clear push to accept the resource, while others feel an unwillingness to defer. From the outside, it looks like priority is given to the one in dire straits, which ethically is often what we strive for. From the inside, each person simply feels what seems reasonable: some feel “I can manage with something else for now,” while one feels “I have to take this chance.” The fairness system, if it exists, is not about giving everyone the same share; it is about giving each what they need to finish neutral. In technical terms, a shared-resource shadow price $\lambda_r(t)$ rises for everyone when a resource is scarce, nudging everyone to consider options with lower $\Delta_r(u)$. Here $\Delta_r(u)$ denotes the amount of shared resource r consumed by option u . Those with short horizons experience a reduced effective penalty because their horizon weighting $\beta(H)$ amplifies compensability relative to resource cost. Here “reduced effective penalty” means the behavioral impact of $\lambda_r(t)$ on choice is moderated by short H in the decision rule; it does not imply the underlying scarcity (or other patients’ needs) is reduced. The result is a quiet, emergent triage consistent with the combined weighting model.

Why do we expect these intensified pressures near the end? The key is that if every life’s ledger $L(T)$ must end at zero, then fewer remaining opportunities imply fewer chances to counteract large tilts. Throughout, “end at zero” should be read as “end approximately neutral” (within the tolerance band $|L(T)| \leq K$ introduced in 6.1.1), not literally $L(T)=0$ in every realization. Mathematically, in a bounded trajectory satisfying $L(T) = 0$, as $T - t$ decreases, the admissible set of future paths narrows. With fewer remaining steps, a large deviation at time t has less opportunity to be offset later. In simple models, the effective weight on compensatory actions increases as remaining time H shrinks, often scaling approximately like $1/H$ under basic assumptions. This $1/H$ form is illustrative rather than required; any monotonically increasing function $\beta(H)$ with $\beta'(H) < 0$ that improves out-of-sample prediction relative to a horizon-free model would satisfy the hypothesis. This is a constraint-driven property of bounded trajectories under a terminal condition. In the Queue System, as the horizon shrinks, the internal stakes on certain actions increase automatically: the system raises the weight on actions that preserve feasibility and lowers the weight on actions that risk irreversible imbalance.

How can we recognize a shrinking-horizon effect in real life? From the inside, people nearing an ending often report a shift in felt tendencies: a kind of brighter guidance toward closure. When time is short, actions that would wrap things up — finishing important tasks, making that phone call, granting or seeking forgiveness — come to mind

more vividly and feel urgently meaningful. There is often an urge toward closure. Conversely, there is heightened friction for irreversible choices that cannot be undone or would be hard to fix later and would leave things unresolved. It is not necessarily fear; it is more like a whole-being sense that “it does not fit in the time I have.” Taking on a drastic new venture or indulging in a risky escape when the end is near just does not align with instincts unless it is something that itself provides closure or fulfillment. Another internal experience is a preference for flexibility: when the horizon is closing, people favor options that keep future choices open, like choosing a reversible course of action over an irreversible one. This makes sense because maintaining flexibility is almost like buying extra time; it keeps the possibility of compensation alive. In summary, near an ending, one’s motivational landscape flattens and narrows in a natural way: grand ambitions fade, interpersonal priorities sharpen, little joys or comforts become significant, and extremes of emotion often level out if good supports are present.

From the outside, an observer can spot several concrete patterns as evidence of the rising shadow price. There is a visible repertoire narrowing: as a deadline or end-of-life approaches, people’s range of activities tends to shrink. They focus on a smaller set of things that truly matter to them, often directly related to final goals or making peace. Interestingly, this does not always mean doing less; sometimes they do more, but in a more focused manner, pouring energy into what counts and dropping the rest. We also see stalling or halting of risky, drawn-out sequences: someone who was considering a long-term gamble might suddenly table it, or a terminally ill person might stop pursuing an arduous legal battle. Long-range projects lose their allure when the long range is not there. And critically, we expect to see affective variance compression — essentially, moods and emotions become more stable and centered as the end nears, provided the person has adequate comfort and support. Empirically, we predict that in the final days or weeks, a person’s day-to-day mood variability is significantly lower than earlier in the life course, even after accounting for other factors. This prediction should be evaluated only when mood measurement remains reliable and after accounting for factors that can mechanically compress variance (e.g., sedation, reduced activity/engagement, and floor/ceiling effects of the rating scale). In plain language, when a person is well cared for near the end, they are often neither euphoric nor despairing, but relatively even-keeled, guided by a gentle emotional landing. We interpret that as the Queue System having applied all possible small corrections to bring the trajectory toward a smooth finish. This is an interpretive hypothesis: variance compression is evidence for the mechanism only insofar as it survives preregistered controls that separate horizon effects from direct stabilization by care, measurement artifacts, and selection effects.

It is useful here to distinguish two kinds of pressure introduced earlier — one from the horizon and one from shared resources. The horizon weight $\beta(H)$ is an internal pressure that ramps up as H decreases; it increases the weight on compensability. When time is short, the menu emphasizes balance-restoring moves and discourages irreversible ones. The shared-resource price $\lambda_r(t)$ for a resource at time t is an external, group-level pressure that rises when many people draw on something limited. A high $\lambda_r(t)$ means that consuming the resource carries a higher effective cost for everyone, because it reduces availability for others. However, a critical design feature is that those with very short horizons may experience a reduced effective penalty, because their horizon weighting $\beta(H)$ amplifies compensability relative to resource cost. In effect, the system can reprioritize near closure without eliminating concern for fairness overall. Thus, urgency and flexible shared-resource allocation coexist: the mechanism is both personal and collective in nature, and inherently compassionate in that it prioritizes those who cannot wait while still asking those who can wait to do so. “Inherently compassionate” is a descriptive shorthand for predicted allocation patterns under $\beta(H)$ and $\lambda_r(t)$ interactions; if short-horizon individuals do not retain differential access under congestion, the horizon-priority hypothesis is falsified.

It is important to note that neither the horizon weight nor the resource price forces anyone’s hand. These factors shape the menu of options that appear salient, but choice remains choice. The admissible set is constrained but not singular — there remains room for individuality and free will. In this way, the promise is that as each horizon closes, a neutral ending can remain possible without turning people into automatons. If a life ends in an irredeemable outcome despite these pressures, that is evidence against the Law, not something the Law can override.

What you’ll get from this Chapter:

- Grasp why balancing intensifies near the end: Understand intuitively that when time left is very limited, there is less opportunity to make up for large ledger imbalances. Debts of suffering or stretches of indulgence become more urgent to resolve. We discuss observable signs of this endgame effect — for example, people urgently seeking closure, making reconciliations, or showing unusual clarity about priorities — and explain why, under a fairness law, these are not just coincidences but expected behaviors.
- Learn the formal notion of a shrinking horizon: See how an agent’s decision policy mathematically changes when the horizon is small. We draw an analogy to economics: as time to use resources runs out, the shadow price on each remaining opportunity rises. A simple model illustrates how the weight on

compensatory actions increases as H shrinks, with $\beta(H)$ behaving approximately like $1/H$ under basic assumptions.

- Explore cross-cultural and end-of-life patterns: Review evidence from hospice care and end-of-life practices across different cultures. We see that terminal patients in very different traditions often exhibit horizon effects, such as forgiving enemies, saying goodbyes, or finding profound meaning in ordinary moments. These observations are treated as candidate signatures rather than confirmation; absence of such patterns in well-supported hospice contexts would count against cross-cultural convergence. We reframe these practices in Queue System terms — reducing irreversibility, increasing flexibility, lowering variance — and discuss how to study them rigorously while maintaining comfort and dignity. We note that comparing experiences across cultures requires assuming a common measurable metric of well-being or variance reduction; if such a metric does not exist or is not comparable, analysis must focus on within-person changes over time rather than cross-group comparisons.
- See the shadow-price equation in action: In a dedicated research note, we formally define the horizon-based weight $\beta(H_t)$ as a function of remaining time H_t and distinguish it from resource shadow prices $\lambda_r(t)$. We then show mathematically how the combined weight on compensability and resource load emerges from a simple weighted utility model. From this formalism, we outline specific predictions — for instance, that certain neural and hormonal urgency signals may correlate with perceived remaining time.
- Identify what to measure for horizon effects: Develop a blueprint of concrete metrics and experimental designs to capture shrinking-horizon dynamics. This includes neural correlates such as control-network activation, physiological markers such as heart rate variability or pupil dilation, and behavioral markers such as changes in choice patterns or response times under time pressure. We discuss how to ethically manipulate perceived horizon in experiments and how to verify that observed changes are due to horizon length rather than general stress or fatigue.
- Recognize fail patterns for horizon scaling: Learn about specific failure conditions that would falsify or weaken the horizon component. For example, if people near an end-of-life event show no special drive toward closure or no unique shifts in behavior relative to matched controls with longer horizons, then the rising shadow price hypothesis is not supported. Likewise, if experimentally shortening perceived horizon yields no difference in choices or feelings in controlled settings, that is evidence against the horizon mechanism. We also distinguish horizon

effects from those driven solely by shared resource strain with no extra horizon factor.

- Understand population-level horizon effects: Extend the horizon principle beyond individuals to groups. We introduce the idea of a population shadow price to describe collective urgency when many people face a short horizon together, such as a community in disaster or a society nearing a known upheaval. We discuss the concept of policy windows — moments when a synchronized rise in horizon pressure can open brief windows for high-leverage, fairness-preserving interventions.

Subsections in this Chapter:

- **6.1 Why Endgame Balancing Intensifies** – Formal intuition and simple proofs that as H shrinks, the admissible set narrows and the effective weight on compensability rises, with $\beta(H)$ increasing approximately like $1/H$ in simple models. We also preview variance compression near closure..
- **6.2 The Intuition of Shrinking Horizons** – Everyday feel of horizon effects, how options drop out or light up as time gets short, and how QS manifests from the inside.
- **6.3 Hospice Across Cultures** – Practical patterns that look like QS in end-of-life care, treated as hypotheses to test rather than proofs.
- **6.4 Research Notes: Shadow Price λ and Horizon H** – Definitions, notation, and estimation: $H_t, \lambda_r(t)$ for shared resources, and the combined weight model $w(u, t) \propto \exp[\beta(H_t) \Phi(u; L, H_t, R) - \sum_r \lambda_r(t) \Delta_r(u)]$.
- **6.6 Fail Patterns for Horizon Scaling** – Pre-registered ways the story could be wrong, such as no $\Phi \times g(H)$ interaction, no menu narrowing near deadlines, or congestion-only accounts fully explaining the data.
- **6.7 Population Shadow Price and Policy Windows** – Extending the idea to groups when many share short horizons, and how synchronized urgency can open brief windows for high-leverage, fairness-preserving interventions.

Where we go next:

We start with the most intuitive consequence of a closing window: when moves are few, the cost of imbalance rises. Section 6.1 explains why endgame balancing should sharpen as horizons shorten; open channels make this signature legible and humane to measure. Channels aid detection; they do not enable the law.

6.1 Why Endgame Balancing Intensifies

As a life (or any bounded episode) approaches its horizon H , the price of non-compensability rises faster than linearly. In simple parameterizations this can appear as a superlinear dependence on remaining time, e.g., through a horizon gain $\beta(H)$ that grows roughly like $1/(H+\delta)$; the precise scaling is an empirical question. Consequently, the admissible menu tilts more sharply toward reparative, reversible, or flexibility-preserving actions the closer one gets to the end. This is not mysticism or “fate” – it emerges as a property of finite path spaces under a neutral-closure constraint.

6.1.1 The path-space argument (one-page proof)

Let $L(t) = \int_0^t F(\tau) d\tau$ be the accumulated net affect (the “ledger”) up to time t . The Law of Fairness demands that at the terminal time T (end of life or episode), the ledger is approximately neutral: $|L(T)| \leq K$, where K is a small allowance relative to typical daily fluctuations. Now consider the remaining horizon at time t : $H_t = T - t$.

If you select some option u at time t , the probability of still finishing neutral will depend on (i) the option’s immediate contribution to compensability ($\Phi(u; L, H, R)$ – does it improve the odds of balancing the ledger or threaten to unbalance it?) and (ii) the number of steps left (the future “coin flips” and choices available) to counteract any damage.

Under broad regularity conditions (bounded affect increments, finite variance processes, etc.), one can show that the set of policies that guarantee a high probability of $|L(T)| \leq K$ must shrink as $H \downarrow$. Here “bounded” means increments of $F(t)$ have finite second moments and no unbounded drift that would trivially violate $|L(T)| \leq K$; if those conditions are not met, the narrowing result does not follow. In intuitive terms: if there are fewer draws left in the journey, you cannot afford as much randomness or risk, because there won’t be enough opportunities later to cancel out any big deviations. Any option that adds a lot of variance or an irreversible drift to the ledger becomes increasingly untenable near the end – those options drop out of the feasible strategy set. Conversely, options that reduce variance, create closure, or preserve future choice flexibility become the only safe bets. In plain language, “*I cannot afford that detour now*” is simply the subjective version of this combinatorial fact.

Two generic theoretical results fall out of this analysis:

- Horizon-weighting grows: The effective weight on the compensability factor Φ must increase as remaining horizon H decreases. In other words, $\beta(H)$ – the coefficient scaling Φ in the decision policy – is an increasing function of $1/H$. Formally, we expect $\beta(H)$ to rise (possibly roughly $\propto 1/H$) as $H \rightarrow 0$. In practice, we implement this with regularization (e.g., $1/(H+\delta)$ or an explicit cap) so $\beta(H)$ stays

finite at the endpoint. This captures the notion that the shorter the path, the more strongly decisions are driven by the need to stay compensable.

- Variance compression near closure: If supportive channels C (for relief, support, etc.) are available, then as $H \rightarrow 0$ the distribution of partial ledgers $\{L(t)\}$ contracts for those individuals. In other words, among people (or days) with only a short time left, we predict seeing much less spread in their ledger values – their trajectories converge to low-variance, closure-focused paths. Practically, this means lives (or projects) nearing their end should look more similar in affect balance (closer to neutral with limited variance) than lives with a long way to go. (*We will treat “compression” as achieving a final ledger variance ≤ 0.80 of a matched mid-life baseline, per our preregistered margin. This 0.80 ratio is a preregistered equivalence margin rather than a universal constant; failure to meet it under adequate measurement and intact support channels counts against the compression claim.*)

These results require no special teleology – they arise from requiring an end condition (balance) on a stochastic process. It’s analogous to how, if you must end a walk at a specific point, your random steps naturally have to get smaller as you approach that point.

6.1.2 Optional stopping intuition (without measure theory)

Another way to understand the endgame effect is by the optional stopping principle. This is a mnemonic for finite-horizon constraints, not a direct invocation of the optional stopping theorem; any formal use would require specifying the process class, filtration, and stopping-time conditions under which the theorem applies. Imagine a stochastic process for affect (the incremental changes in L) where at each moment you have some control $u(t)$ to choose (from your admissible set) that can influence the drift of that process. If you are far from the end, you can tolerate exploratory moves that might occasionally worsen the ledger, because you have many future steps to potentially repair any damage. However, if the process will stop soon (i.e. H is small), you lose those future repair opportunities; thus the only safe strategy is to choose controls that ensure the expected contribution to $|L(T)|$ remains small. In effect, as the stopping time nears, your policy must become *risk-averse in a very specific way* – not risk-aversion to *reward*, but aversion to any action that could drive the ledger out of the neutral bound.

This is exactly what the Φ function encodes: how much an option u improves the chances that the process ends within the neutral band. If the horizon is short, only options with $\Phi \approx$ positive (balancing) will maintain the high probability of neutrality, whereas options with Φ negative (imbalancing) become prohibitive.

A compact, practical form of the decision weight that aligns with how we model data is:

$$\omega(u, t) \propto \exp[\beta(H) \cdot \Phi(u; L, H, R) - \sum_r \lambda_r(t) \cdot \Delta_r(u)], \text{ with } \beta(H) = \beta_0 + \beta_1 / (H + \delta).$$

Here we've made an example specific functional form for $\beta(H)$: $\beta(H) = \beta_0 + \beta_1 / (H + \delta)$, where δ is a small constant to prevent β from literally blowing up to infinity at $H = 0$. This form captures the idea of a smooth but accelerating tilt: as H gets very small, $\beta(H)$ gets large, but not infinitely large – there's a cap to how strongly it can weight things at the final moment, reflecting that people don't become completely single-minded robots at the end, but they do narrow their focus significantly.

This weighted-choice formula is a quantitative translation of common sense: *when you're almost out of time, the only options that "feel right" are those that won't wreck your chance of closure*. Empirically, the acceleration claim should be treated as a hypothesis until it is supported by preregistered longitudinal data with clear horizon instruments and appropriate confound controls. And empirically, we do observe something like an accelerating tilt near real-world deadlines and clinical endgames – behavior changes gradually at first, then markedly in the final stretch.

6.1.3 Irreversibility, flexibility, and “synthetic time”

Three key properties of actions determine how strongly the endgame horizon constrains them:

- **Irreversibility $I(u)$.** If doing u creates a state that is hard or impossible to undo (burning a bridge, making a permanent change, an invasive action with no return), then as the horizon shortens, the effective “drift toward terminal risk” caused by u rises. In simple terms, irreversible actions become the first to drop off the menu as $H \rightarrow 0$, because there is no room to recover from them. They carry a hidden high cost that grows with $1/H$.
- **Flexibility $Flex(u)$.** Conversely, actions that are easily reversible or are exploratory “probes” act like synthetic time. They keep more future branches open, effectively extending how long you have to compensate. Near the end, the sensitivity of Φ to flexibility is much higher. We could say $\partial\Phi/\partial Flex$ increases as H decreases. This is why in endgame situations we see so many trial balloons, tentative outreach efforts, and small reversible steps – people instinctively favor moves that *behave as if they had more time*.
- **Noise control (variance reduction).** Any option that lowers the variance of future affect (for instance, getting a good night's sleep, taking pain medication, staying in a stable routine) becomes disproportionately attractive when time is short. Reducing noise is akin to shrinking the error bars on the remaining trajectory,

which is exactly what you need if you have to thread the needle to neutral. So, near endgames, actions that stabilize the situation (even if they don't directly add pleasure or remove pain) gain priority.

A practical parameterization for modeling these effects in experiments or data might break Φ into components:

$$\Phi(u; L, H, R) \approx w_1 \text{ReliefGain}(u) + w_2 \text{RepairGain}(u) - w_3 I(u) + w_4 \text{Flex}(u) - \sum_r \lambda_r \Delta_r(u).$$

Here λ_r denotes the shadow price $\lambda_r(t)$ at the decision time (index suppressed), and $\Delta_r(u)$ is measured in the corresponding resource units. Here *ReliefGain* and *RepairGain* capture how much u immediately helps the ledger (either by providing relief from suffering or repairing a negative), $I(u)$ is irreversibility, $\text{Flex}(u)$ is flexibility, and the last term subtracts resource draw costs as before. Importantly, the weights w_3 and w_4 on irreversibility and flexibility should themselves increase as $H \downarrow$. In practice, one can model that by including interaction terms like $I \times (1/H)$ and $\text{Flex} \times (1/H)$ in a regression – effectively letting the model learn that irreversibility matters a lot more when horizons are short, etc.

6.1.4 Multi-agent pressure: horizon vs. congestion

As discussed, two prices interact in a multi-person setting:

- Horizon price $\beta(H)$: This is an individual, private pressure that escalates as H shrinks. It tilts a single person's choices toward compensability (it's "internal urgency").
- Congestion price $\lambda_r(t)$: This is a social or systemic pressure that escalates as a shared resource r gets loaded. It's the ambient push for everyone to conserve a scarce channel.

One key consequence of having both is a predicted horizon-priority effect under congestion: In a crowded system, yes, everyone's menus co-move toward low-draw substitutes as $\lambda_r \uparrow$ (think of everyone in a busy clinic trying to use less of the nurse's time). However, those with short horizons will be partially exempt from this penalty: their high-need options remain available to them because $\beta(H)$ overrides some of the social penalty for them. Mechanistically, this corresponds to an $H \times \lambda$ interaction implemented at the group or policy level (others defer, or triage rules allocate capacity); it does not imply that scarce capacity is created or that congestion vanishes. This isn't favoritism, it's feasibility – if they didn't get priority, they might not reach closure at all, violating the fairness constraint. In observable terms, this means if we look at a congested scenario, we expect to see menus "tilting together" for the group (everyone becomes more frugal

in using the resource), *but* the people with urgent needs still somehow manage to get served (e.g. short-horizon patients still get that ICU bed or that emergency slot).

This horizon-priority is something we can test for: it's a kind of interaction between individual state and group state. Later we formalize it as a term κ that effectively reduces λ for short-horizon individuals (Section 6.4.3). If horizon priority did *not* occur, that would correspond to a fail pattern ("Congestion-Only World") wherein everyone just experiences the same resource shortage effect and the truly urgent cases *don't* get reliably through – which would be a mark against the Law of Fairness being a real guiding principle.

6.1.5 Neural and interoceptive consequences

If endgame balancing is real, it should leave footprints in the brain and body. These are proposed correlational signatures to test, not claims that any single region is uniquely necessary or sufficient for the effect. As noted (and cf. Section 5.3 for the underlying neuroscience), we expect systematic modulation in known circuits:

- **vmPFC/OFC:** We should observe a stronger *value signal* for high- Φ options as H decreases. For example, in fMRI, the ventromedial prefrontal cortex might show increased activation to a "repair" choice (say, helping someone or resolving a problem) especially when one's time is nearly up, compared to earlier on. Such activation patterns would be interpreted as correlational signatures within a preregistered GLM; they do not by themselves establish causal enforcement of the constraint. This corresponds to $\gamma_\Phi > 0$ and increasing with shorter horizon in a GLM (see Section 6.5.1 for specific predictions).
- **rIFG/STN:** These inhibitory control regions should show heightened activity (or effective connectivity) for braking low- Φ actions under short horizons. In a lab task, this could manifest as longer stop-signal reaction times or stronger beta-band power in EEG when trying to suppress a temptation while on a short deadline. Depending on operationalization, "braking" can appear as proactive slowing (longer go RT in stop contexts) and/or improved reactive stopping (higher stop success at matched delays, often corresponding to shorter SSRT under the standard estimator); the preregistered task definition should fix directionality. Neural models predict a significant $\Phi \times (1/H)$ interaction here – essentially an increased "braking" effect as H is smaller.
- **ACC:** The anterior cingulate should encode the cost of continuing a plan that might lead to imbalance when time is short, as well as the social congestion cost. So ACC might have two roles: one as part of the horizon mechanism (detecting "this action will make it impossible to finish neutral given the time left") and one

as part of the λ mechanism (noticing “the environment is crowded, this is costly for everyone”). For instance, ACC signals might ramp up when a person considers a risky move during a short-horizon period, reflecting an internal “Don’t do it” warning proportional to how little time remains. Similarly, ACC might show activation correlating with the hospital ward being full, etc., integrating those cost signals.

- Anterior Insula and Autonomic System: We expect a somatic marker pattern where the body effectively says “yes” to high- Φ moves and “no” to low- Φ ones, more emphatically as the horizon shrinks. This could be seen as changes in heart rate variability (HRV), skin conductance (SCL), or pupil diameter. For example, just imagining doing the “right thing” (a closure action) might produce a calming parasympathetic response when time is short, whereas imagining a pointless detour might produce a subtle stress response (“heaviness”). The insula, which integrates interoceptive signals, would likely be part of that awareness – one might literally feel the difference in one’s gut.

(Research in Section 6.5 will detail how to measure these neural and bodily signatures in practice.)

6.1.6 Diary, telemetry, and field signatures

How do shrinking horizons appear in real-world data? We can think of personal diaries or other observational datasets. Empirically, as a horizon contracts, we expect to see:

- Repertoire narrowing: The person initiates fewer distinct types of actions per day. Instead of doing ten different things, they might do just three, focusing on those with closure value. The proportion of their activities that directly contribute to “wrapping things up” increases. (For example, in a digital trace, one might see fewer unique apps or conversation topics, and more of them related to final goals.)
- Stickiness asymmetry: Once high- Φ behaviors (like writing to loved ones, finishing tasks, seeking help) are started, the person tends to carry them to completion. In contrast, low- Φ behaviors (idle distractions, unnecessary errands) are more likely to be cut short or aborted mid-stream. For instance, you might start drafting an email to reconcile with someone and definitely send it (near end-of-life), but you might start scrolling social media and then stop after a minute because it doesn’t “hold” your attention like it used to.
- Contact graph tilt: Socially, the pattern of interactions shifts toward repair and support contacts and away from adversarial or trivial ones. A person nearing an ending will, for instance, call close family or old friends more, and engage less in

debates or business networking. Their communication network “tilts” toward sources of relief or meaning.

- Affect variance compression: Day-to-day swings in net affect (mood) get smaller for those whose channels of support are intact. For example, in the last week of a semester, a student’s mood might be more even (even if under pressure) because they’re singularly focused – whereas during mid-term, they had big ups and downs. Statistically, among people with objectively short horizons (final semester, final month of life, etc.), the variance in daily affect should be lower than in matched individuals who are not near an ending (provided the short-horizon group has access to support; if not, we might see distress instead – an important caveat).

All these signatures are things we can measure with modern tools: smartphone sensors and apps can log activity diversity, communication patterns, mobility, etc. Experience sampling can capture mood variance. Importantly, we would preregister indices for each (e.g. a “repertoire index” of unique actions per day) to avoid cherry-picking. In the analysis, we’d model counts of actions or contacts as Poisson variables (checking for overdispersion >1.2 and using a Negative Binomial with log link if needed) to properly quantify repertoire changes. Many of these measures can be gathered without intruding on the person – for example, by passively logging phone use or via brief diary entries – which is crucial in sensitive contexts like end-of-life.

6.1.7 Concrete vignettes (minimal drama, maximal testability)

To build intuition and also suggest study designs, here are some low-drama, common scenarios recast as testable vignettes:

- Final week of a product release: Take a software engineer, Alex, with a list of tasks. Ten days before release, Alex’s mind still wanders to “maybe we could refactor this subsystem” or other exploratory ideas (moderate-risk detours). Two days before release, those thoughts don’t even come up – or if they do, they have no grip. Instead, Alex finds it easy to concentrate on writing final tests and documentation (closure tasks). If we instrument this scenario, we’d see menu thinning (fewer new ideas considered), ease asymmetry (quick engagement with wrap-up tasks, resistance to tangents), and a team-level effect: as the build server or testing environment becomes a bottleneck (shared resource congestion), *all* engineers shift to lighter usage of it, except the ones working on release-critical fixes, who still manage to monopolize it without contest. This would illustrate horizon-priority under a shared constraint.

- Hospice reconciliation: A man named John has avoided calling his estranged father for months. Earlier in the year, the thought of calling was easy to put off (many other things felt equally pressing or the emotional cost felt high relative to benefit). Now the father is in his final week of life. Suddenly, for John, calling becomes the obvious and only thing to do – it almost “does itself,” whereas any alternative activity feels pointless or wrong. The family around them might have, months ago, tolerated the stalemate, but now everyone encourages the call and creates a quiet space for it. This vignette shows the potent effect of a truly short horizon on overcoming long-standing inertia. It’s not that John consciously calculated the change; the horizon effect manifests as a felt shift in priorities. Measuring this could be done by narrative coding: looking at timing of reconciliatory acts in hospice patients/families relative to prognosis updates, etc.
- University finals: A group of college students gradually cease social outings and streaming binges as finals approach. By the week of exams, most of their leisure time is channeled into group study sessions or simply sleeping (to preserve functioning) – behaviors directly serving the goal of passing exams with minimal pain. Interestingly, students whose exams finish earlier (so their horizon for academic tasks is shorter sooner) display this convergence earlier than those who have later exams. This creates a natural gradient to observe: we’d predict a measurable $\beta(H)$ effect where, say, library study-hour counts or social media usage differ across students in proportion to how soon their next exam is. This is a benign setting to test horizon theory: all students face the same overall workload, but their perceived horizon for needing to be “done” is staggered.

These vignettes emphasize *testability*. They avoid dramatic one-off stories and instead suggest patterns we can observe in ordinary behavior given different horizon lengths. Each could be turned into a study: instrumented product teams, hospice communication logs, student time-use diaries, etc., to quantitatively capture the signals of horizon effects.

6.1.8 Discriminating from stress, fatigue, or risk aversion

It is crucial to distinguish shrinking-horizon effects from other phenomena like stress, generalized fatigue, or simple risk aversion. It’s easy to misattribute changes to the wrong cause if we don’t carefully separate them. The Queue System hypothesis makes specific predictions that differentiate horizon effects from these look-alikes:

- Selectivity vs. shutdown: Stress or fatigue often cause a broad shutdown – you do less of *everything*, good or bad, because you’re exhausted or overwhelmed. Horizon effects, by contrast, are selective: they don’t blunt *all* activity, only the

low- Φ , low-compensability activities. High- Φ actions may actually increase even if they require effort. For example, a stressed person might stop socializing entirely, whereas a short-horizon person might socialize *more* if those interactions offer closure or comfort, while dropping other trivial activities. We look for this selective pattern (repair actions up, frivolous actions down), which stress alone wouldn't produce (stress would likely drop both).

- Reversibility preference (vs. low-effort preference): When people are simply tired, they often choose whatever is easy in the moment (which might be indulging or procrastinating – *not* necessarily what's long-term beneficial). In horizon-driven behavior, the preference is not merely for low effort, but specifically for reversible or option-preserving actions. A horizon effect individual might actually undertake something effortful (like an emotionally intense conversation) if it's the right kind of action (high Φ and often reversible in steps) instead of an easy indulgence. So it's not just "do easier things" – it's "do things that keep the future open or repair the past," which might actually require effort.
- Instrument sensitivity (arousal independence): A key test is using a credible deadline prime or other horizon manipulation that *does not increase arousal*. If we can change someone's perceived horizon (e.g. make a deadline salient) in a way that doesn't significantly raise their stress (no big change in heart rate, pupil, etc.), yet we see their choices shift toward closure-oriented ones, that's a clear horizon effect. Stress and fatigue, on the other hand, usually come with arousal changes (cortisol spikes, etc.). So we can include physiological covariates: if the behavior changes remain even after controlling for those (or if our manipulation kept arousal flat), we know it's not just stress. For example, instructing someone that "actually, this task will end sooner than expected" might alter their decisions while their pupil size remains the same – indicating a cognitive horizon effect, not a fight-or-flight response.

In short, horizon effects should show specificity: a targeted tilt in the decision profile, rather than a generic reduction in activity or a blanket aversion to risk. If we find that everything a person does just diminishes equally near the end, or that only tasks with high effort drop out but not the indulgences, then perhaps we are observing fatigue or rational energy budgeting, not the fairness-driven horizon mechanism. We have to design studies to pull these apart (e.g. include measures of stress, include both high-effort indulgences and low-effort repairs as options, etc.). If after all that, the distinctive signs (like preference for reversible moves, persistence of effort in the right places, and arousal-independent choice shifts) are not present, then the "endgame balancing" story is weak.

(Fail Pattern 6.6.5, the “Non-Selective Narrowing,” specifically addresses this scenario of generic shutdown vs. selective tilt.)

6.1.9 Minimal math for practitioners

For quick application (e.g., by a practitioner running an experiment), we can summarize the decision model in a simpler logistic form. One useful representation is:

$$\text{Pr}(\text{choose } u) = \text{softmax}[\theta_U U(u) + \theta_C C(u) + \theta_\Phi \Phi(u) + \theta_H \Phi(u) H^{-1} - \sum_r \theta_{\{\lambda_r\}} \lambda_r \Delta_r(u)].$$

Here the horizon interaction term is $\theta_H \cdot \Phi(u) \cdot H^{-1}$, and the minus sign begins a separate resource-cost term $-\sum_r \theta_{\{\lambda_r\}} \lambda_r \Delta_r(u)$. This equation is a more accessible way to plug in factors. In words: the probability of choosing option u is given by a softmax (logistic selection) where the predictors include:

- $U(u)$ = utility or immediate reward of u (with coefficient θ_U),
- $C(u)$ = generic conflict or cost of u (with coefficient θ_C),
- $\Phi(u)$ = our compensability score for u (coefficient θ_Φ),
- $\Phi(u) \cdot H^{-1}$ = the interaction of compensability with the inverse horizon (coefficient θ_H , representing the horizon gain effect),
- $\sum_r \lambda_r \Delta_r(u)$ = the resource draw of u times shadow prices (with coefficients $\theta_{\{\lambda_r\}}$ for each resource type).

The predictions to preregister from this model are clear:

- $\theta_H > 0$: there should be a positive coefficient on the $\Phi \times (1/H)$ term. This means choices indeed reflect a stronger preference for high- Φ as H gets smaller (horizon scaling effect). If our fits find $\theta_H \approx 0$ consistently, that’s evidence against intensification.
- $\theta_{\{\lambda_r\}} > 0$: positive coefficients on the resource cost terms. This means congestion matters – using up a scarce resource deters choice, all else equal (social penalty effect). We’d expect, for example, $\theta_{\lambda_t} \text{ time} > 0$ if time-with-nurse is scarce: options needing more nurse time get penalized.
- Overall model comparison: A model including the horizon term and λ terms should predict choices better (especially out-of-sample) than one with just utility, conflict, etc. We would plan to demonstrate an improvement in predictive metrics (like cross-validated log-likelihood or WAIC) when θ_H and θ_λ terms are included. If adding these terms yields no predictive gain beyond standard factors, then the law isn’t earning its keep as a model component.

(In practice, we will report something like: did adding Φ and $\Phi \times H^{-1}$ improve WAIC or LOO by a significant amount over a baseline model? – using one such metric in this chapter to avoid redundancy, per guidelines.)

6.1.10 Fail Pattern: What would count against intensification

We conclude the theory section by explicitly stating what empirical findings would count against the hypothesis that endgame balancing intensifies as described:

- Fail Pattern – Flat-Slope Null: $\Phi \times H^{-1}$ interaction is found in behavior or brain data when immediate utilities are controlled and horizons are credibly manipulated. Correction: the Flat-Slope Null is that the $\Phi \times H^{-1}$ interaction is not found (i.e., is indistinguishable from zero under those controls). In other words, people show no extra bias toward compensatory actions even when time is short; their choices look the same as if H were long. This would directly contradict the core claim.
- Fail Pattern – No Endgame Signature: No repertoire narrowing or stickiness asymmetry appears as natural deadlines approach, despite using high-quality measurements. If we carefully track people during, say, the final week of a semester or the last weeks of life (with consent and support), and we see that they continue to have just as broad a menu and abort rate for meaningful actions as anytime before, then the theory loses a key prediction.
- Fail Pattern – Congestion-Only World: In group scenarios, we see congestion effects (co-moving menus under load) but no horizon priority. If in a crowded resource situation every individual behaves as if they only respond to λ and those with short horizons don't get any special consideration (their outcomes are no better than anyone else's), then the horizon component of the theory is in trouble. It would imply maybe it's all just rational allocation of resources with no additional "must finish" push.
- Fail Pattern – Rival Models Cover It: If a simpler or fundamentally different model (like a homeostatic set-point model or a reinforcement learning model with some generic cost for waiting) can reproduce all these effects without needing Φ or horizon terms, and it fits the data as well or better, then our law isn't necessary. For instance, if a resource-rational RL model with a clever penalty structure can mimic repertoire narrowing and final variance reduction, and adding our law-specific terms doesn't improve predictive accuracy, that's a serious strike (Rival Sufficiency scenario).

Any replicated combination of these outcomes – observed with adequate power, controlling for confounds – would force us to downgrade the status of endgame balancing from a deterministic constraint to maybe just a tendency or heuristic, or even

abandon it altogether as a special phenomenon. Science-wise, that is an acceptable outcome; it would mean fairness isn't "baked in" as strongly as we thought, and lives might often end unfairly unless other forces intervene. But until such results appear, the hypothesis stands that a shrinking horizon inherently intensifies the push toward balance.

6.1.11 Where we go next:

Intensification is only meaningful if people can feel and show it. 6.2 traces the inside view (what tightening balance feels like) and the outside view (what others can observe) so we can separate horizon effects from fatigue, stress, or wishful readings.

6.2 The Intuition of Shrinking Horizons

You already know this feeling in everyday life. A deadline suddenly becomes real. A goodbye moves from hypothetical to definite. A diagnosis turns the abstract idea of limited time into a concrete timeline. Without anyone explicitly telling you what to do, your inner landscape changes. Certain options lose their appeal or “stickiness.” Others – like apologizing, getting something important done, reaching out, or simply resting to conserve energy – become easier than they were just yesterday. This section gives crisp language to that ordinary experience of a shrinking horizon and shows how to tell it apart from more familiar states like stress or fatigue.

6.2.1 What it feels like from the inside

From the first-person perspective, a shortening horizon reveals itself through several subtle shifts:

- Menu thinning: You can still *think* of all the usual detours or idle diversions you might take, but they don’t grip you the way they used to. Your mind keeps returning, almost unbidden, to the actions that would “close the loop.” It’s as if the array of tempting distractions has physically thinned out – they appear in your thoughts, but translucent, easy to ignore.
- Ease asymmetry: The actions that would repair, resolve, or complete something feel *surprisingly light*. It’s easy to get started on them, and once started, you have energy to finish. In contrast, actions that would escalate a conflict, or that carry long-term consequences you can’t fix, feel *heavy*. It’s a palpable effort to even muster interest in them.
- Somatic nudge: You experience a bodily intuition – a “yes” feeling for the next right move, and a “no” feeling for a diversion. For example, imagining writing that apology email might come with a sense of relief or “rightness” in the gut, whereas opening a video game or picking a fight yields a subtle stomach tension or heart sink. These are the somatic markers aligning with high-Φ vs. low-Φ options.
- Temporal reframing: You become acutely aware of time and keeping options open. You start to naturally frame choices in terms of “If I do this now, will I still have time to adjust later?” Reversible actions (send a tentative message, schedule a short meeting) feel valuable because they preserve future choices; you sense that doing something irreversible now is like burning precious time. In short, reversibility feels like borrowed time.

A handy shorthand someone gave is: a shrinking horizon converts an “I could do anything” mindset into an “I can keep *these*” mindset. You stop feeling you can do

everything on your list (or anything at all), and instead feel “I have to be careful and choose among these few actions I can definitely carry through.”

6.2.2 What it looks like from the outside

To an observer (or in third-person data), the same situation would be reflected in concrete changes in behavior and outcomes:

- Repertoire narrowing: The person engages in fewer distinct activities each day, and more of those activities are directly aimed at closure or stability. For example, if we categorize someone’s daily activities, as a deadline nears, categories like “administrative chores” or “random web browsing” might disappear, leaving mostly “work on main task,” “coordinate with team,” “sleep.” The *diversity* of actions drops, while the *fraction of closure-related acts* rises.
- Stickiness asymmetry: Once a person initiates something beneficial (like a task that has high Φ), they’re likely to finish it—especially as the horizon shrinks. Meanwhile, if they start a low- Φ diversion, you often see them stop halfway or quickly lose interest. Imagine a student: a week before exams, they might start watching a movie and then turn it off after 10 minutes because it just doesn’t feel right to continue. But if they start reviewing notes, they’ll go until done. Earlier in the term, the opposite might have been true (easy to binge, hard to study). This asymmetric completion rate is a telltale sign.
- Contact graph tilt: Their pattern of social contacts shifts. They initiate more communications with people who matter for closure (family, close friends, mentors) and fewer with casual acquaintances or adversaries. If you mapped their email or messaging connections, you’d see a convergence toward “inner circle” connections and supportive roles, with a drop-off in frivolous or conflictual exchanges.
- Variance compression: If you chart their daily mood or net affect, you’d see the swings shrinking as the end approaches *provided key needs are met*. For example, those nearing project completion with sufficient support might show more stable mood day-to-day than earlier in the project. In end-of-life contexts, patients in hospice often show more emotional balance (less extreme despair or elation) compared to earlier, *if* they have pain managed and support in place. Technically, we’d measure a reduction in the variance of daily affect ratings for short-horizon individuals relative to a baseline or a control group not at an ending. (This is one of the signals we aim to quantify, remembering to check that things like open support channels “C” are present – since without them, short-horizon might instead mean continued distress.)

These signatures are not just anecdotal; they are things we can track in data. Diaries, phone logs, wearable sensors – all can capture elements of this outside view. For instance, a smartphone telemetry study might find that in the week before an exam, students' GPS data shows fewer distinct locations visited (repertoire narrowing), their app usage shifts to more educational apps and less social media (closure tilt), and sentiment analysis of their texts might show more messages to family or study partners and fewer in trivial group chats (contact graph tilt). Additionally, nightly mood check-ins might show reduced volatility in stress ratings as the exam nears (if they have adequate support like good sleep etc.). All this would paint a clear picture of horizon effects in action.

(In later chapters, especially Chapter 13 on telemetry, we discuss how to gather such multi-channel data ethically and what it tells us.)

6.2.3 The cues that trigger the horizon effect

What causes a person's mind to conclude "the horizon is shrinking now"? There are a few distinct cues that can generate this state, often working together:

- Horizon cues (time signals): These are external or internal signals that *time is objectively short*. Examples: a date on a calendar (like exam date approaching), a doctor saying "likely only a month left," a manager announcing "48 hours till we lockdown features." Even something like boarding a plane for a one-time opportunity can be a horizon cue ("after this flight it's now-or-never"). These cues feed directly into the perception of H_t . A credible deadline or endpoint signal is usually necessary to kick the horizon effect into gear.
- Ledger cues (balance signals): This is the felt sense of how imbalanced or incomplete things are (L state). For instance, a person might have a nagging feeling "I owe someone an apology" or "I really need a day of rest." That feeling raises the salience of certain actions that would address it (repairs or relief). In our model, it effectively raises the utility of repair actions. If a person feels they have a big unresolved pain (debt) or unmet need, as the horizon shrinks this ledger cue amplifies the push to resolve it immediately.
- Resource cues (shared-channel signals): These tell the person about the state of shared resources around them. Seeing a queue or congestion ("lines are getting long," "everyone is busy," "the clinic is full") signals that λ is high. This tilts the menu toward using alternatives or lower-draw options. However, an important nuance: if someone also has a short horizon, that horizon cue can override resource caution (they'll still pursue the high-resource action if it's critical). But absent a short horizon, resource cues alone can cause a temporary change in

behavior that might look like horizon behavior (people act more frugally with resources). We need to disentangle these; often they co-occur (e.g., near end-of-life there's often both urgency and limited resources).

The same person can have different horizon perceptions in different contexts of their life. For example, a scientist might feel a shrinking horizon for an upcoming grant deadline (work context) but simultaneously feel an expanding horizon in personal life after resolving a big issue (home context). The mind tracks these context-specific horizons separately. So a cue in one domain (like a work deadline) will strongly affect behavior in that domain without necessarily spilling over, unless the domains interact (e.g., both work and personal life drawing on the same pool of mental energy, which they do to an extent).

6.2.4 How to tell horizon effects from stress or fatigue

We touched on this in Section 6.1.8 conceptually; here we list the concrete discriminators one can look for or experiment on:

- Selectivity vs. shutdown: If we observe a person under high workload or pressure, is their behavior selectively changing (some things drop, others increase) or uniformly dropping? Shrinking-horizon predicts selectivity (they still put effort into certain actions), whereas general stress predicts a more uniform drop (they just reduce effort across the board). For example, a fatigued person might procrastinate on *everything*, while a horizon-effect person might procrastinate on unimportant things but surprisingly *increase* effort on crucial ones.
- Reversibility preference: One can design a choice experiment where some options are equal in difficulty but differ in reversibility (e.g., an easy indulgence that's irreversible vs. an equally easy indulgence that's reversible later). Under fatigue, a person might take either as long as it's easy; under a horizon mindset, they'll prefer the reversible one. Similarly, compare a *difficult reparative action* vs. a *difficult indulgent action* – a stressed person might avoid both (too difficult), whereas a horizon-driven person would tackle the reparative one despite difficulty but avoid the indulgent one.
- Arousal-controlled horizon primes: We can actually test this by manipulating horizon without changing stress. For instance, have participants do a task with a countdown timer that either indicates plenty of time or very little time, but ensure the task difficulty is the same and measure physiological arousal. If the *short timer* condition alters their strategy toward more cautious or compensatory moves even if their heart rate and pupil dilation stay the same as in the long timer condition, then we have isolated a horizon effect. Stress effects almost always

correlate with some arousal (pupil dilation, sweaty palms, etc.). We would also include a condition where we increase stress (like time pressure with alarms, etc.) but maybe keep horizon long, to see if that pattern differs. The unique fingerprint of horizon effect is behavior change without corresponding stress markers (or with stress markers accounted for in analysis).

In empirical studies we plan, we will incorporate these checks. For example, we'll have pupillometry or HRV monitors on participants when giving them sudden deadline cues. If performance or choices shift but pupil/HRV does not, we attribute it to horizon-specific cognitive factors. If adding pupil and HRV as covariates wipes out the effect, then maybe it was just stress after all (*Fail Pattern: Arousal Substitution* would be considered).

In summary, horizon effects are *selective, strategic, and not solely driven by physiological stress*, whereas plain stress is indiscriminate and broad in its impact (everything suffers) and fatigue is about doing less of anything effortful (regardless of its future value). These differences are not just academic; they determine how we intervene. If someone's behavior change is due to stress, the remedy is to relieve stress. If it's due to a horizon effect, the "remedy" (if any) might be to ensure they have the right options available (since their mind is doing something adaptive given the constraint).

6.2.5 Micro-phenomenology: a one-minute self-check

As a brief experiential exercise (one that readers or study participants can do to recognize the horizon effect in themselves):

1. Name one unfinished thing that truly matters to you. It could be a relationship that needs mending, a project you deeply care about, or a personal goal left hanging.
2. Ask yourself: "What is the smallest *reversible* move I could do *right now* that would advance closure on this?" Maybe it's sending a text message, writing a single paragraph, or setting up a short meeting – something that moves things forward but doesn't lock you in fully (you can adjust if needed).
3. Imagine doing that action, vividly, and notice your body's response. Is there a sense of relief, a gentle ease, perhaps a drop in tension when you visualize actually doing it? That feeling – if it's there – is the signature of a high-Φ option when your mind is considering a shorter horizon.
4. Now, contrast it: think of a tempting but irreversible detour you could also do (like binge-watching a show or sending an angry reply, or making a big impulse purchase). Imagine indulging in it right now and note how your body feels. Is there a subtle heaviness, a pit in the stomach, or a flutter of anxiety? That is the brake –

your interoceptive system flagging a low-Φ option under (perhaps unconsciously) a tighter horizon mindset.

5. Repeat this mental exercise for a few days in a row. Most people can quickly learn to recognize the difference between the “horizon closing” guidance and mere stress. They report that by day 3 or so, they *know* the feeling of a horizon-aligned action (it comes with a peculiar clarity or calm energy) versus a mere distraction (which feels hollow or uneasy).

This exercise is anecdotal but aligns with what our theory suggests. It’s essentially a guided way to feel the Queue System’s tug. Interestingly, it doesn’t take an actual terminal diagnosis to experience this – even setting an arbitrary personal deadline can induce a mild version of it, which supports the idea that our brains are constantly doing some version of this calculus at different scales.

6.2.6 Field vignettes (low drama, high resolution)

To ground this further, here are some realistic vignettes one might collect in field studies, illustrating horizon effects with everyday characters:

- The unsent reply: On Monday, Alice composed a scathing reply to a colleague’s annoying email. She hasn’t sent it yet. By Thursday night, with a critical joint presentation on Friday morning (a near horizon for their collaboration), that harsh draft no longer “holds up” in her mind. The thought of sending it feels heavy and unproductive. Instead, a neutral or conciliatory reply, or even letting it go, now seems far more natural. The draft stays unSENT. This illustrates how as a shared deadline looms, the menu shifted – conflict escalation lost its appeal without anyone explicitly advising Alice; it’s the horizon making an unwise action feel unwieldy.
- The pre-flight call: Ben has been ambivalent for weeks about calling an old friend he had a falling out with. There was no urgency, so he kept putting it off. The night before Ben flies overseas for a year (a clear horizon on the opportunity to reconcile), he dials without ceremony and has the conversation. Nothing externally forced this change; *only the horizon drew nearer*. The fact that no new argument or event occurred, except the impending departure, underlines that it was the horizon effect in pure form – a nearer horizon made the choice clear.
- Studio release day: A music producer and her band have spent months tweaking an album. One month out from the release date, they tinker endlessly – new ideas keep coming (horizon was still somewhat open). On the eve of release, all those tweaks lose attractiveness. The band members collectively feel “it’s done enough.” They freeze the tracks and get a good night’s sleep instead of pulling an

all-nighter to make last-minute changes. This is an example in a creative context where, as horizon shrank to zero, the mind favored closure and rest to ensure a stable release, whereas before, exploration was psychologically supported.

- Ward at capacity: In a palliative care ward, as bed occupancy hits 100%, staff begin naturally steering incoming stable patients to alternative care (e.g. home hospice) – everyone's choices co-move toward preserving resources. However, one patient with rapidly declining condition (short horizon) still “keeps” the last infusion slot and gets full in-hospital care, and interestingly all staff and other patients intuitively support it. They might say it's obviously appropriate. In analysis, that's horizon-priority: even though the ward is full (high λ for beds), the system (people's collective behavior) effectively exempts the patient who truly cannot wait. No formal rule needed to be invoked aloud; it happens through the felt imperative that “this person needs it now.” This vignette ties back to our earlier theoretical scenario and shows how it plays out in practice.

Each vignette pairs a *before* (longer horizon scenario) with an *after* (short horizon scenario for the same person) and highlights measurable behaviors: unsent vs. sent emails, call made vs. not, decisions to stop vs. continue tweaking, allocation of resources. These could be turned into case studies or qualitative data in field research. The power is that they show the horizon effect isolating itself – the friend's conflict was longstanding but got resolved only when horizon changed, implying horizon was the key variable.

Collecting a series of such vignettes across contexts (work, personal, medical) with timestamps and details provides rich evidence that the phenomenon is general and not tied to any one type of goal or stressor. We prefer low-drama examples because they are common enough to gather systematically (not everyone has a melodramatic reconciliation, but many have small-scale versions of these things).

6.2.7 Where we go next:

If shrinking horizons are real in lived experience, they should leave cultural footprints. 6.3 looks across traditions and care systems for convergent “horizon engineering”—practices that quieten prices, safeguard dignity, and preserve feasible compensation.

6.3 Hospice Across Cultures

End-of-life is the sharpest example of a shrinking horizon. It's where the horizon H becomes explicit and unavoidable. Remarkably, around the world, communities have independently evolved practices that look very much like they are managing horizon effects: they clear channels, reduce distractions, and focus on reconciliation, relief, and meaning in the final weeks. In other words, without using our theoretical vocabulary, many hospice and terminal-care traditions behave as if they were deliberately reducing irreversibility, increasing flexibility, lowering variance, and protecting short-horizon access to scarce resources. This section describes those recurring patterns across cultures, reframes them in our LoF/QS terms, and then discusses how to measure these effects and what would count as evidence against a universal horizon effect at end-of-life.

6.3.1 Recurring practices that resemble horizon engineering

Across diverse settings – from modern hospital hospices to home-based palliative care, from monastic infirmaries to indigenous healing practices – several motifs repeat worldwide:

- Reconciliation windows: It's common to arrange visits or mediated conversations for patients to forgive, apologize, or say goodbye. For example, families will call estranged members to visit, or clergy/mediators will be brought in to facilitate forgiveness. QS interpretation: This raises Φ by yielding a big *RepairGain* (mending relationships) often at low resource draw (short visits, letters). Essentially, communities create a space for high- Φ , low- λ interactions to happen.
- Legacy projects: Many traditions encourage the dying to create an ethical will, record messages, sort photos, or narrate their life story. QS view: These turn diffuse regrets into concrete closure tasks with high OptionFlex. Recording a message or making a memory book is done in small reversible steps (you can pause, edit, etc.). It helps integrate one's story (reducing irreversibility of unspoken things by giving them voice).
- Symptom relief as capacity creation: Controlling pain, shortness of breath, anxiety, and ensuring sleep are universal hospice tenets. QS: These interventions reduce variance and provide *ReliefGain*, stabilizing the ledger and the patient's capacity to engage. In our terms, effective analgesia or anxiolysis directly lowers the “noise” in affect and increases the feasibility of positive acts (it's hard to reconcile if you're in agony; relieve the agony and Φ for reconciliation goes up).
- Time structuring: Many hospice programs implement quiet hours, simplified daily routines, and predictable schedules. QS: This lowers cognitive load (“noise”) and

ensures that key high-Φ actions remain keepable. If a patient knows every day there's a quiet afternoon, they can plan to have an important talk then. Structured time is like giving them control and reducing random shocks, which preserves compensability.

- Social triage: Explicit prioritization of access to scarce services for those near end-of-life. E.g., a hospice might guarantee that if a patient is in their final days, a nurse or counselor will be there (even if it means others wait). QS: This is horizon-priority in resource allocation: short-horizon streams get precedence on shared channels (transport, clinician time, special ceremonies).
- Ritualized forgiveness/absolution: Many religions have specific end-of-life rituals for confession, forgiveness of sins, or community reconciliation circles. QS: These are formal mechanisms to achieve repair with minimal logistics. They often require just a conversation with a priest or a simple ceremony – low resource, high impact on Φ (guilt or regret is lifted, ledger moves toward neutral).
- Use of calming sensory inputs (touch, music, scent): Hospices often have volunteers for hand-holding, soft music, aromatherapy or familiar scents. QS: These help modulate the interoceptive signals – shifting the body into a calmer state where *high-Φ options “feel right” and low-Φ detours “feel wrong.”* For instance, gentle touch and lullabies might reduce fear and make acceptance or reconciliation feel safer. It's essentially tuning the somatic markers to favor closure (insular signals of safety).
- Night vigil / watch: In many traditions, someone stays with the dying through the night, or family takes turns (vigil). QS: This guards the low-draw counterweights like sleep and dreaming – ensuring the person isn't alone or panicked at night. It reduces nocturnal spikes of anxiety (no “3am terrors” if someone is there to comfort immediately), preserving flexibility for next-day choices. Also, continuous presence can help catch pain or distress early (quick relief -> stability).
- Narrative compression (life review): Activities like life review therapy, story circles, or reciting meaningful narratives. QS: These integrate fragmented experiences, helping the person see a coherent story (which lowers psychological conflict variance) and allow them to express unsaid truths in a reversible form (words, which are more flexible than unresolved actions). In effect, it reduces irreversibility of secrets or unresolved feelings by *voicing* them in a safe context.

These motifs appear in Catholic last rites, Buddhist final rites, Hindu and Muslim end-of-life customs, Jewish vidui (confessional) and tahara (purification), Indigenous practices focusing on harmony, and even secular hospice protocols. The surface forms differ (prayers vs. letters vs. circles), but they all converge on what looks like *horizon*

management. Different cultures found different *implementations*, but they seem to solve the same control problem: how to maximize the chances of a peaceful, balanced closure when time is almost up.

This convergent evolution suggests that *neutral closure* is indeed a recognized (if implicit) objective and that human societies have intuitively grasped some aspects of the fairness constraint. It's as if culture provides the "macro" Queue System when needed: a set of customs that make sure the conditions for fairness are met as the horizon closes. We can frame culture as a library of control policies (some emphasize one aspect, some another) that all share the same goal: allow a life to close without heavy regrets or unresolved pain.

Interestingly, many of these end-of-life practices have everyday parallels. Cultures have long used smaller-scale rituals – from weekly days of rest to regular fasting or confession – to prevent our ledgers from spiraling in normal times (see Chapter 21.5 on spiritual practices as "queue hygiene"). In essence, the same principles that guide a peaceful final closure are echoed in routines that keep life balanced all along.

(*We should note: not everyone gets access to these ideal practices, and outcomes vary. But the existence of these norms is telling.*)

6.3.2 Cultural variations as parameter tuning, not different physics

While the motifs are shared, cultures vary in how they achieve them. We can think of these differences as tuning parameters rather than completely different laws:

- Who leads reconciliation: In some cultures a priest or imam plays the key role; in others a family elder or therapist might. The function (raising RepairGain via reconciliation) is the same, but the *agent* differs. *QS perspective*: It doesn't matter who does it as long as someone does – the parameter is "who is the trusted agent for facilitating repair." This doesn't change the underlying dynamic; it's like using a different catalyst for the same reaction.
- Individual vs. family agency: Some cultures give the dying person individual autonomy to decide their priorities; others operate via family consensus (the family decides how to allocate time, who visits when, etc.). *QS*: The unit of coordination differs (individual vs. group policy), but the objective is the same – minimize irreversibility, maximize closure. One culture might reduce conflict by having family shield the patient from certain info, another by ensuring the patient's wishes are followed exactly. Either way, they aim to avoid chaotic, irreversible outcomes.

- Affect tone of the setting: Compare a solemn Catholic hospice with a celebratory African home-going tradition. One is quiet and serious, another might be filled with songs and family gatherings. QS: Both can work because each *stabilizes physiology and meaning* in its own way. A familiar solemnity can calm some; a familiar celebratory vibe calms others. Parameter: affective tone that best reduces anxiety for that culture. Either approach reduces the variance and conflict if it's what people expect and find fitting.
- Symbolic resources used: Different faiths or communities use different symbolic acts – water, fire, beads, incense, prayer cloths, etc. All these are typically low-cost cues that provide comfort or mark a transition, helping the person and family process what's happening. These symbols can reduce fear, resolve internal conflicts (someone may let go once a particular prayer is said), and make reparative or accepting choices “stickier” because they feel meaningful. Think of it as various culture-specific ways to achieve a positive somatic marker or reinforce a decision.

So, one culture might look superficially very different from another in end-of-life practices, but those differences are like using different colored threads to weave the same pattern. Culture = control policy library: each culture offers a set of parameter settings that have proven effective historically at meeting the final fairness constraint (or at least at aiming for a dignified closure).

An important implication: when studying horizon effects across cultures, we should not assume the *absence* of a behavior (say, open verbal forgiveness) means absence of the effect – it might be achieved differently (e.g., through a ritual rather than direct apology). We need to look at functional equivalences. This means our measures should be somewhat culture-adapted (configural invariance) but seek the same core patterns.

(From a Law of Fairness standpoint, we'd say nature imposes the constraint (neutral closure) and cultures have, over time, found ways to respect that constraint in humane and meaningful ways, even if they didn't frame it as such.)

6.3.3 Hospice as a system that lowers shadow prices

In our theoretical language, a well-run hospice or end-of-life care program does three crucial things:

1. Increase horizon gain $\beta(H)$ on the right things: It makes actions that provide relief or repair extremely easy to do (low psychological and logistical barriers). For example, hospices often have on-call mediators or spiritual counselors (“It’s not too late to say...” scripts) and encourage patients to voice needs. These amplify

- the value of high- Φ options. Essentially, the system is boosting Φ or effectively making $\beta(H)$ larger for those actions – guiding the person to see closure opportunities as salient and doable.
2. Lower social penalties $\lambda_r(t)$ on critical channels: The system protects access to resources needed for end-of-life tasks. It might reserve transportation for last wishes, allocate clinician time for symptom control or important family meetings, enforce quiet hours so that rest (a key resource) is available. By doing so, it reduces λ for those channels for the dying person (and their family). That means using those resources doesn't feel costly or greedy; it's guilt-free and expected. Everyone else adjusts around that (the community or hospital accepts some inconvenience to keep those channels open for short-horizon needs).
 3. Add synthetic time via flexibility: Good programs encourage small, reversible gestures – e.g., “You can just try a short visit” or “Write a draft letter, you don't have to send it,” or they facilitate short calls rather than long formal meetings. By doing this, they increase the *Flex* part of actions and effectively give the person more perceived time. The idea is to keep options open even in uncertainty. If a patient is unsure about reconciling, hospice might suggest a brief, informal meeting rather than a dramatic confrontation – so it's easily abortable if it goes poorly. This *flexibility* means the person doesn't feel it's “now or never” in a do-or-die way, ironically giving them the confidence to proceed.

When those three conditions are met, families often report outcomes like “the right words found us” or “things fell into place at the end”. In our view, nothing mystical occurred – rather, the environment was engineered such that the menu tilted and stayed open long enough for all the critical closure actions to happen. Essentially, hospice created conditions for the Queue System to operate optimally: it kept barriers low for good options, kept bad options largely out of the picture, and allowed the natural balancing process to play out.

This reframing of hospice practice in QS terms is useful because it suggests testable metrics and improvements. If a hospice intervention isn't working well, we might diagnose: Is it failing to provide flexibility? Are we still imposing too high a λ (e.g., bureaucratic hurdles that consume time)? Are we not amplifying the value of meaningful acts (maybe we need a chaplain or someone to encourage those conversations)? The model provides a kind of checklist for evaluating end-of-life care from a systems perspective.

(It also underlines ethically that these practices are not just compassion – they might be necessity if the law is true. “Protected windows for reconciliation” aren’t just nice-to-haves; they would be the means by which fairness is achieved at life’s end.)

6.3.4 Cross-cultural field signatures to measure

If our account is right, we should observe certain regularities across cultures regardless of doctrine or specific ritual, as long as the structural conditions (open channels, horizon awareness) are present. Here are outcomes we’d predict in any culture’s end-of-life setting (assuming no major external impediments):

- Repertoire narrowing toward closure: In the final weeks, the fraction of a person’s actions that fall into categories like reconciliation, legacy-making, relief-seeking, or rest should increase, while trivial or diversional acts drop away. This pattern should appear *across cultures*, provided those channels exist. For example, whether it’s writing letters (Western) or performing specific rituals (maybe more common elsewhere), the person’s activities converge on meaning-making and comfort-related acts.
- Stickiness asymmetry: Once patients or family initiate a repair or comfort behavior (visiting an old friend, engaging in a religious ritual, accepting help), they are likely to carry it through completely; conversely, indulgent or non-essential acts (planning a future vacation, arguing about minor issues) often get dropped mid-course near the end. We expect to see, e.g., higher completion rates for forgiving phone calls, but many half-finished attempts at, say, watching random TV or idle chat, as the end nears.
- Horizon-priority in resource use: If resources are scarce (they often are), short-horizon patients will end up receiving a greater share of scarce, high-value resources (like emergency medical attention, or the limited time of certain experts, or access to sacred rites at peak times) *with the community’s endorsement*. We could measure this by looking at, say, how hospital bed usage or pain medication allocation correlates with prognosis: does the distribution show that those nearer death get relatively more? If the system is fair, it should. And qualitatively, other patients/families should not object strongly – because implicitly they understand the priority (if they do object, it may indicate a communication or policy problem).
- Variance compression in daily affect: Among those who have proper symptom control (“intact channels”), we expect to see a shrinkage of day-to-day mood variance as they approach the final days. Their emotional highs and lows moderate – perhaps because big swings get corrected by care (if upset, someone

comforts them; if in pain, medication given; if too elated or unrealistic, someone grounds them). In places where these channels are missing (e.g., poor pain control), you might *not* see compression – you might even see *increased* variance or just persistently high distress. Also, if someone’s horizon isn’t acknowledged (they are kept in the dark about their condition), they might not show these patterns either. But in properly run hospice, we predict a notable convergence to moderate affect. We could test this by mood diaries or retrospective family reports, and importantly compare to a matched group not in hospice.

- Dream counterweights: This one is interesting and culturally less noted but we predict: even if external channels are limited, dreams might provide a compensatory outlet. There are many anecdotes of patients having vivid reconciliation or relief dreams (like dreaming of a deceased loved one forgiving them, or reliving a happy memory) and then the next day they act with more peace or reach out to someone. We suspect an increase in such dream content (e.g., more reports of meaningful dreams) and that when such dreams occur, there’s a measurable uptick in positive behavior next day. If a culture emphasizes dreams, this might be more documented (some indigenous cultures do place weight on final dreams). But even if not, an attentive study might find patients reporting, “I had a dream about X and now I feel I can let go of that issue.”

All of these can be studied with surprisingly lightweight methods: short daily checklists, simple logs from caregivers, interviews, etc. Importantly, these measurements can and should be done without interfering with care or violating dignity. For example, a daily checklist filled by a nurse or family might note categories of patient activity (reconciliations attempted, legacy work done, social interactions, etc.), mood (rough rating), symptom levels, dreams recalled (if any), etc. These can be aggregated to see the trends described. Because this is sensitive, all participation must be consent-based and designed so it doesn’t burden or distress the patient or family.

(We’ll soon discuss in 6.3.5 and 6.3.6 how to design such studies ethically and what minimal intrusion protocols look like.)

One key point: any cross-cultural measurement must account for measurement invariance. We must ensure that, say, what counts as “reconciliation behavior” is defined comparably across contexts, or use relative measures (like each person’s change from baseline). We might need configural invariance at least – the concept exists in each culture even if manifest differently. If strict invariance (same numerical values) isn’t possible due to cultural expression differences, we rely on within-person and within-culture comparisons primarily. This is where our earlier mention of metric invariance

comes in: we assume at least that we can rank-order experiences similarly across groups, or we restrict claims to qualitative pattern presence/absence rather than exact magnitudes, to be safe.

6.3.5 Minimal-intrusion research designs

Studying end-of-life phenomena demands utmost sensitivity and ethics. We cannot treat dying people as guinea pigs; any research must *benefit* them or at least not burden them. Here are components of a study design that could achieve the insights without violating dignity or comfort:

- Common-core protocol: Develop a short, neutral daily checklist that can be gently administered (by a caregiver or self-reported if the patient is able). It might include simple yes/no or 1-5 questions: “Did you sleep well? Pain under control? What social contacts did you have today (family, friends, clergy)? Did you complete any small tasks (letters, calls)? Any regrets softened? Any dreams?” This should be translated and culturally adapted for each setting so it’s in familiar language and tone. It should take just a few minutes. The neutrality is key – it should not push any agenda or make implications about what *should* happen, just record *what is*.
- Shared-channel meter: A simple log of wait times or availability for key services: e.g., “How long did it take to get a nurse when needed today?” or “Were you able to have private time with your family when you wanted it?” or “Did you experience any delay in getting pain medication?” These are objective-ish measures of resource access (could be kept by staff). We can z-score these and combine them to form an index of congestion λ_t conditions: high wait times, many denials, etc., mean high λ . Conversely, some measure of slack (like hours of quiet time provided, or spare capacity) indicates low λ . This lets us correlate patient behavior/mood with resource pressure context.
- Horizon index: We can record both a clinical estimate (the doctor’s best guess of time left, updated periodically) and a subjective question like “Does it feel to you like you have: days / weeks / months / more than months left?” (some cultures might be uncomfortable asking the patient this; if so, could skip patient’s own estimate). This gives us a perceived vs. objective horizon measure. If direct questions are too blunt, one could use projective ones like “Are you living more one day at a time or still planning months ahead?” The goal is to gauge H_t from both perspectives.
- Outcome markers: Things like completion of legacy tasks (did they finish that photo album?), number of reconciliations achieved (perhaps ask family if there

were important connections made), observed calm or distress (nurse's rating of how peaceful the patient appeared), caregiver burden change (did family feel more at peace or more strained), and a follow-up with family after death about their satisfaction or regrets. These outcomes help see if the horizon management succeeded (not just in theory but in leaving people feeling that closure was achieved, to the extent possible).

- Ethics note: This must be opt-in only. Patients and families should explicitly consent and be told clearly that participating or not will *not* affect their care. They can withdraw at any time with no consequences. The questions or measurements must be crafted in a culturally appropriate way, likely vetted by local clinicians and possibly spiritual counselors to ensure they're not offensive or upsetting. For instance, some cultures might not want to ask the patient about their own horizon; then we don't. Also, the data collected should be kept confidential, used for improving care or knowledge, and ideally even fed back in some helpful way (e.g., if a patient expresses unresolved issue in a survey, the care team addresses it – thus the research directly benefits care). *Most importantly: participation must never ever reduce the care itself.* Research personnel should operate in the background or be part of the care team, not demanding extra work from caregivers or hogging time that should be spent with the patient.

In practice, one might embed a researcher or trained volunteer within the hospice team who does these observations as part of routine. If anything, it could even improve care by systematically noting if certain best practices happened each day.

This careful design ensures we gain insights responsibly. We can learn, for example, how strongly various practices correlate with outcomes, or which fail patterns (from 6.3.8 below) might be occurring in some places, thereby guiding improvements.

(It's worth noting: purely observational studies like this, with consent, are generally ethical and often approved by families who want to help others in the future. The key is sensitivity and transparency.)

6.3.6 Vignettes across traditions (function over form)

To further illustrate the cross-cultural sameness-in-function, let's sketch a few vignettes from different traditions, emphasizing the function each practice serves in QS terms:

- Theravāda monastery infirmary: In a Buddhist monastery, a dying monk is reciting metta (loving-kindness meditation) daily; novices ensure strict quiet during certain hours (no disturbances). Families come for short, structured visits (often guided by a senior monk). Function: Reduce variance and conflict – the

environment is serene (lowers noise), and reconciliation is facilitated in brief, meaningful encounters (high RepairGain with minimal strain). Flexibility is preserved (visits are short but frequent rather than one long ordeal). The focus on balance/beauty (*hózhó* in a Navajo context, but similar aim here of harmony) integrates narrative and gives the dying monk peace that things are in order.

- Catholic hospice: A patient receives Anointing of the Sick (a sacrament for forgiveness/peace). A social worker or chaplain systematically helps them make any needed apologies or farewells (sometimes via letters or calls). Function: This converts what could be large irreversibilities (sins or regrets) into symbolically reversed form – the anointing absolves, the apologies spoken address guilt. It also lowers λ by ensuring a chaplain's time is reserved for them (explicitly part of care). Pain is managed with medication promptly (variance reduction). Essentially, the sacraments and support staff cover the reconciliation and relief bases, respectively.
- Muslim family home: The extended family rotates so someone is always there reciting the Ya Sin (a calming, relevant chapter of the Quran) for the patient; neighbors bring food (reducing burden on immediate family); a local imam comes to guide forgiveness prayers. Function: Community support reduces caregiver λ (they don't burn out – more capacity for meaningful interactions), the rhythmic recitation stabilizes interoception and provides spiritual comfort (lower anxiety, consistent "background" that reduces cognitive noise), and the imam's role raises RepairGain by religiously framing forgiveness. It's a different setting (home vs. hospital), but does the same fundamental things.
- Navajo (Diné) setting: Emphasis on *hózhó* – maintaining balance and beauty. A hataalii (traditional healer) performs a chantway ceremony aiming to restore harmony. Family and community ensure certain taboos are observed (to avoid conflict with beliefs). Function: The chant integrates the narrative ("your life is part of a harmony"), reducing existential conflict and giving a sense of closure. It also, through ritual, reduces irreversibility of any spiritual disharmony by ritually *fixing* it. Social conflict is minimized because everyone participates in a unifying ritual – any interpersonal tensions are often sublimated in that shared spiritual goal. Essentially, it's an alternate route to the same end: the person and their circle feel "in balance" which is their term for neutral ledger.

These vignettes show different "flavors" but the same control logic. Each has:

- A way to lower noise (be it silence, prayer, familiar routine).
- A way to amplify repairs (religious confession, guided visits, ceremonies that implicitly reconcile).

- Steps to preserve flexibility (short visits, reversible gestures like symbolic acts rather than forcing big real-world changes).
- A recognition of priority (the dying person is the center of attention; normal rules like who eats first, who gets resources, are adjusted).

By documenting such cases, we can reassure ourselves that our model isn't just Western-centric but captures something universal. It also prepares us to measure it appropriately: for example, in a Navajo study, "reconciliation acts" might be recorded as "participated in chantway" rather than explicit apologies, but functionally that is their reconciliation mechanism.

6.3.7 Practical guidance for programs

Before moving to the fail conditions, it's worth summarizing some practical tips for hospice and palliative programs that emerge from this understanding (many programs do these already, but framed in QS terms):

- Protect the counterweights: Ensure the basics like sleep, pain control, and calm are *always* prioritized. In QS language, guard those low-draw, high-impact channels. For instance, enforce quiet hours at night, aggressively treat pain and breathlessness, and give the patient periods where nothing is demanded of them (so they can dream, reflect, or just recuperate). These actions stabilize the menu so the person has the capacity for closure acts.
- Script small reversible steps: Don't push giant emotional confrontations or bucket-list marathons. Instead, offer *templates and gentle openings* – e.g., suggest they write a short note rather than a long letter, or have a brief phone call instead of an in-person if that's easier, or meet one person at a time instead of a huge gathering. Provide example phrases to break the ice ("I just wanted to say..."). This makes potentially overwhelming tasks feel doable and safe (high Flex). The patient can always do more later if they want, but this gets things started without fear.
- Make horizon information clear and kind: If appropriate in the culture, be honest with patients (and families) about the prognosis in a compassionate way. Knowing roughly how short time is can *activate the horizon effect* positively – often patients themselves suspect it. Euphemisms or avoidance might delay the beneficial shifts. On the other hand, being clear ("weeks rather than months") but coupling it with reassurance ("we will support you to accomplish what you need") can empower patients. Essentially, give the *H_t* signal in a kind way; obfuscation might delay their natural prioritization and cause regret later.

- Publish triage rules: If not already explicit, let everyone know that, for example, “In this hospice, when resources conflict, those nearer end-of-life get priority for critical resources.” That transparency can reduce resentment or confusion, and actually ease the minds of patients (longer-horizon ones won’t feel neglected at random; they’ll understand why someone else is getting more attention now, and they can expect the same when their time comes). It lowers social λ in a sense that it reduces conflict over resources.
- Integrate cultural leaders: Bring in the appropriate cultural or religious figures if the patient/family desires – they are often low-draw amplifiers of comfort and meaning. A priest, imam, elder, or respected community member can achieve huge RepairGain or OptionFlex with just a conversation or a simple ritual, which staff might not achieve on their own. It’s an efficient way to boost Φ (they often know exactly what to say or do to relieve burdens in that cultural context).
- Measure what matters (and iterate): Keep track of those “common-core” markers we discussed. For instance, note how many reconciliations happened, or how often patients seem to sleep well, etc. Use these to adjust resources. If you notice, say, many patients have unresolved dreams or regrets, maybe allocate more counseling. If family conflict is stopping closure, bring in mediators earlier. Essentially, treat it like a learning system: observe menu co-movements (like many families complaining about lack of private time could mean that channel is congested – maybe institute new policy to create that space).

By following such guidelines, hospice programs can effectively implement the Law of Fairness without ever calling it that – simply by creating an environment where balancing can naturally occur. And these guidelines are evidence-based (or at least theory-based) rather than just tradition: they directly map to the levers we know are crucial (H , Φ , λ , Flex, etc.).

(Now we’ll shift to considering how our broad claims could be disproven by cross-cultural data – essentially, what would indicate that not all cultures follow these patterns or that something else is at play.)

6.3.8 Fail pattern: what would count against the cross-cultural claim

To fairly test our cross-cultural horizon hypothesis, we must admit what findings would seriously challenge it:

- Null repertoire narrowing in supportive contexts: If we find a culture or setting where, despite open channels and credible awareness of dying, patients do *not* narrow their activities toward closure – e.g., they continue random or adversarial

activities at the same rate as before – that's a red flag. Especially if in multiple cultures with good hospice practices we still see no narrowing, it could mean our presumption of universality is wrong.

- No horizon-priority in resource use when scarcity is real: If, across sites, when resources are scarce, those near end-of-life *do not* receive any prioritization or don't even seek more (and everyone treats them just the same as others, or they themselves refuse needed resources equally), that would undermine the idea of an internal or systemic horizon effect. It could mean fairness isn't baked in as strongly as we think.
- Equal stickiness of low- Φ and high- Φ acts near the end: If data showed that near end-of-life, people abandon high-value tasks just as often as low-value ones (i.e., no stickiness asymmetry – they drop meaningful conversations mid-way just as much as they drop TV shows), then our theory fails to describe a real phenomenon. It would suggest perhaps people are just shutting down uniformly or behaving randomly.
- No variance compression in affect even with symptoms managed: If even in well-supported end-of-life cases, we see affect variance *not* decreasing (or even increasing), then one of our key quantitative predictions fails. We'd have to consider that maybe variance compression is not a necessary outcome (perhaps adaptation, not fairness, is at play or something else).
- No link between reconciliation/relief dreams and actions: If our speculation about dreams is tested and fails – say, lots of short-horizon folks report reconciliation dreams but it has zero bearing on their next-day behavior, or no increase in such dreams is observed at all – then the dream angle might be bogus. It wouldn't kill the whole theory but would remove one proposed compensatory channel.

If we observed consistent nulls across multiple sites and cultures on these points, it would seriously undermine our claim that hospice practices worldwide converge on horizon-aware balancing. We'd have to entertain that maybe what we're seeing is culturally constructed and not a human-universal response to horizon. Perhaps some cultures encourage fairness at end-of-life, and others do not (and maybe fairness isn't a law then, just a culturally specific value).

So far, anecdotal evidence leans in our favor, but robust data could surprise us. Thus, we list these fail outcomes as what to watch for.

(From an investigative lens: if, say, we see null repertoire narrowing in one culture, we should also check that our measurements were valid for that culture. It might be that

(reconciliation took a form we didn't count. So a true falsification should come after careful verification that we measured properly – akin to ensuring metric invariance as mentioned. If truly no pattern exists even when appropriately measured, then yes, that's a falsification.)

If a strong pattern of these nulls emerged, it would undermine the cross-cultural convergence hypothesis and suggest that what we think is a “law” might be more malleable by cultural or individual differences than expected. We might then revise the Law of Fairness to a narrower scope or incorporate additional variables (like maybe cultural norms can override the balancing tendency in some cases).

6.3.9 A one-page field guide (for clinicians and chaplains)

Summarizing the actionable insights of this chapter so far, here's a quick field checklist that any clinician, chaplain, or caregiver might use at end-of-life (implicitly drawn from our fairness framework):

- Name the horizon gently and concretely. Ensure the person and family have a shared understanding of the prognosis in clear terms. This cues the natural prioritization process, though it must be done with compassion.
- Open low-draw channels first. Make sure basic comfort measures are in place: effective pain relief, a quiet environment, unhurried time. These channels (sleep, relief, calm) are foundational; they keep the person capable of engaging in higher-order closure tasks.
- Offer small, reversible scripts for repair and legacy. Don't just tell someone “you should reconcile”; enable it. E.g., “You could start with a short call – even five minutes – just to say hi,” or provide materials like blank cards to write notes, or suggest “Is there a story or message you'd like to record? I can help with that.” Lower the activation energy for these acts.
- Reserve scarce resources transparently for short-horizon needs. Whether it's the last ICU bed, or simply the main recliner in the living area – if someone is in their final days, make it clear they get priority. Others will adapt if it's explained as fairness (most people intuitively get it). It prevents resentment and actually often others become cooperative allies when they know why.
- Invite cultural/spiritual rituals that stabilize physiology and meaning. Encourage whatever practices are meaningful to the patient – be it prayer, meditation, singing, or even a last outing if possible. These often act as state stabilizers and provide a sense of completion or peace (which is basically boosting Φ or lowering psychological noise).

- Notice the tilt: Pay attention to when a patient shows inclination toward a “right action” (e.g., suddenly expressing desire to contact someone) – then remove any obstacles quickly (help them get in touch, ensure privacy, etc.). Similarly, if they show disinterest or aversion to something that used to engage them (maybe a hobby or TV), don’t force it for the sake of “keeping them occupied” – recognize that as a sign their menu is tilting and respect it. In short, go with the grain of their horizon-driven preferences.
- Record simple markers (repairs, rests, dreams) to guide tomorrow’s plan. Debrief at the end of each day or shift: did they get good rest? Any unresolved pain or anxiety episodes? Did they achieve any connection they wanted? Use that to adjust the next day – e.g., if they were too tired all day, tomorrow prioritize energy management; if they kept bringing up a person they miss but didn’t call, facilitate that call next day.

This “field guide” isn’t labeled in scientific terms, but it’s directly informed by them. It helps ensure that the care team is effectively acting as the Queue System’s ally – maintaining conditions for fairness to play out.

Bottom line (for Section 6.3): Cultures around the world, knowingly or not, have built systems to handle life’s endgame in a way that *looks like* the Law of Fairness in action. By studying and learning from these practices, we both validate our theory and improve care. And if we ever find a society or scenario that routinely violates these patterns and yet people feel at peace, then we’ve truly learned something new that could challenge this law.

6.3.10 Where we go next:

Human stories point to parameters. 6.4 formalizes them—defining and estimating horizon H and shadow prices λ —so we can stop hand-waving and start modeling the joint signals we expect to see in behavior and brain.

6.4 Research Notes: Shadow Price λ_t and Horizon H_t

This section is a practical notebook for estimating and experimentally manipulating the two key “prices” that govern admissible sets: the horizon gain $\beta(H)$ associated with time left, and the shared-resource shadow price λ_t associated with congestion. It’s meant as a reference for building preregistered models, designing instruments, and creating measurement pipelines that can withstand adversarial scrutiny. We break down definitions, share methods for estimation, and highlight pitfalls.

6.4.1 Definitions at a glance

First, let’s clearly define our terms and notation in one place:

- **Horizon H_t :** This is the effective time remaining (from time t to the end T) in which compensatory moves can happen. It’s not always just chronological time; it can integrate multiple factors like prognosis, deadlines, reversibility, expected delays, etc. If T is the known (or estimated) end, then in simplest form $H_t = T - t$. But effective horizon might be shorter if some things can’t be done in parallel or if certain windows close early.
- **Shadow price $\lambda_{r,t}$:** This is the marginal penalty for using one unit of shared resource r at time t . For readability we write this as $\lambda_r(t)$ or $\lambda_{\{r,t\}}$ throughout; any doubled subscript here is just “resource r at time t ,” not a second time index. We consider various types of resources: could be a caregiver’s time, a limited number of slots (like a dialysis machine availability), attention of the team, or even intangibles like “quiet atmosphere.” Each such resource r has a congestion level. $\lambda_{r,t}$ converts that into a cost added to any option u that consumes $\Delta_r(u)$ units of resource r .

We often talk of a combined “social penalty” for an option u at time t as:

$$\text{Penalty}(u, t) = \sum_r \lambda_{r,t} \Delta_r(u)$$

where $\Delta_r(u)$ is the amount of resource r that option u would use. (For example, $\Delta_{\text{bed}}(\text{"admit patient"}) = 1 \text{ bed}$, $\Delta_{\text{nurse}}(\text{"hour of special nursing care"}) = 1 \text{ hour}$, etc.)

- **Choice weight model:** We can write the *instantaneous* weight $\omega(u,t)$ for an option u (suppressing the individual index if focusing on one person) as:

$$\omega(u, t) \propto \exp[\beta(H_t) \Phi(u; L_t, H_t, R_t) - \sum_r \lambda_{r,t} \Delta_r(u)]$$

This condenses what we had earlier. The notation R_t in $\Phi(u; L_t, H_t, R_t)$ indicates that Φ may also depend on the context of resource availability (e.g., if the patient knows something is in short supply, that can affect perceived compensability). In practice, we

treat $\beta(H)$ and λ as separate factors. We also note $\beta'(H) < 0$ in typical scenarios, meaning as H decreases, $\beta(H)$ increases (its derivative with respect to H is negative). This formalizes horizon urgency. With these definitions set, our job as researchers is to estimate H_t and $\lambda_{r,t}$ from data and to design experiments that manipulate H or λ cleanly.

6.4.2 Estimating H_t : objective and perceived horizons

We usually need to quantify “how much time is left” in two senses: an objective horizon (according to best external evidence) and a perceived horizon (according to the person’s own mind). We model H_t as a latent state that can be informed by both.

A. Objective horizon (clinical/organizational):

- *Clinical horizon:* In medical contexts, one can use survival analysis models (Weibull, Cox proportional hazards, etc.) incorporating known prognostic factors (labs, symptoms, functional status) to get an expected remaining time. For instance, $\hat{H}_t^{\{\text{obj}\}} = \mathbb{E}[T - t | X_t]$, where X_t is the set of covariates at time t (like current health metrics). This gives something like “estimated 3 months” or “90th percentile of survival is 2 weeks,” etc. This is a continuous update as new data comes in (disease progresses, etc.).
- *Operational horizon:* In project or task contexts, if T is a deadline, one might adjust it for expected waiting times or rework. Example: if deadline is in 10 days but you anticipate you’ll spend 2 days waiting on feedback and likely 1 day redoing something, then effective time to finish tasks is 7 days. Formula might be: $\hat{H}_t^{\{\text{obj}\}} \approx (T_{\text{deadline}} - t) - \mathbb{E}[W_t] - \mathbb{E}[R_t]$, where W is wait time and R is rework time expected. Tools like PERT charts or critical path analyses can inform these estimates for horizon in projects.

These objective metrics give a baseline H . They often come with uncertainty (like a distribution of possible $T - t$). But they provide an anchor.

B. Perceived horizon (psychological):

- *Meters:* We can directly ask people, or use standardized questionnaires. For example, ecological momentary assessment (EMA) item: “It feels like I have ___ left to do what I need to” with choices like “a short time / some time / a long time.” Or scales like Zimbardo’s Time Perspective Inventory, which includes a Future subscale. Shorter validated scales also exist for time perspective or future orientation. We can deploy a single-item repeated measure in daily assessments: “Do you feel your time is running out? (1–5)” etc.

- *Latent state fusion:* Ideally, we combine objective and subjective. We can set up a state-space model where H_t is a hidden state that drifts downward over time and is informed by multiple observations. For example:

$$H_t = H_{t-1} - \Delta t + \epsilon_t, y_t^{(k)} = a_k H_t + \eta_t^{(k)}, (k = 1..K)$$

where the first equation says horizon naturally counts down (with some noise ϵ if unpredictability), and the second is multiple indicators y (like perhaps a clinical prognosis-based estimate, a self-report, maybe a caregiver estimate), each linearly related to H_t with some coefficient a_k and noise. We'd run a Kalman filter or particle filter to estimate H_t over time from these signals, yielding a best guess and uncertainty.

In simpler terms, we treat H as a dynamic latent variable updated as new info comes. By reporting posterior bands for H_t , we acknowledge uncertainty.

- *Instrument checks:* If we manipulate perceived horizon via an experiment (like giving a deadline prime), we must verify that we changed perceived H without confounding arousal. For example, if we show a countdown timer to simulate urgency, we should measure pupil or heart rate to ensure this didn't inadvertently spike stress; if it did, we must include those as covariates or adjust the design (maybe a more subtle prime).

So in our data, we'll often have a sequence of H -hat values for each person over time. We might even define something like "horizon ratio = subjective H / objective H " to see if someone feels shorter or longer horizon than would be expected – which could be interesting (some people are optimistic, some pessimistic about time left).

C. Instruments to shift H_t without changing utility:

Experimentally, how can we alter someone's sense of horizon *without* altering the inherent value of options or raising generic stress?

- Credible deadline primes: E.g., in a lab task, one can publicly commit to an earlier end time than participants expect, or display a countdown clock that signals time nearly up. It's crucial that the tasks' rewards don't change – only the time to complete them. For instance, if you want to test horizon in a game, suddenly announce "we're ending in 2 minutes, wrap up what you can" (some participants) vs. "you have plenty of time" (others). This has to be credible (they must believe it). Also, often a between-subject or within-subject with counterbalancing design helps.

- Milestone revelation: Provide information that some future opportunity is the “last one” or that a phase of the task is nearly over. E.g., in an educational context, reveal a hidden grading rule like “anything not done by this interim checkpoint can’t be made up later,” which effectively creates a closer horizon for some goals. If participants weren’t aware of a constraint then become aware, their perceived horizon for certain actions shrinks.
- Synthetic time grants: The opposite – extend horizon artificially by adding reversible checkpoints. E.g., allow “draft” submissions or “practice rounds” which essentially increase effective H by boosting flexibility (like Section 6.1.3 said). In an experiment, one group might be told “you can revise your answers later” (giving them a sense of more future opportunities) vs. another told “all answers are final” (strict horizon). This tests flexibility as synthetic time.

(In manipulations, always ensure any observed changes are not just because one group got more value – e.g., telling someone they can revise might also reduce their anxiety or change their motivation structure. We try to isolate the effect: ideally both groups have equal incentives and difficulty, just different horizon framing.)

- Checks for confounds: Whenever we do such things, we measure arousal proxies (pupil, skin conductance, heart rate variability) to confirm we didn’t unintentionally cause stress (which could mimic horizon effects). If arousal does change, we treat it as a covariate or confound to address.

6.4.3 Estimating $\lambda_{\{r,t\}}$: from queueing to shadow prices

A. From waits to prices: In many service systems, you don’t directly observe an abstract “cost” for using a resource—you observe delays or queues. Queueing theory links such delays to implied cost. Consider a simple M/M/1 queue (Poisson arrivals, exponential service, one server):

$$E[W_q] = \rho / [\mu (1 - \rho)] \text{ where } \rho = \lambda_{\text{arr}} / \mu \text{ is utilization.}$$

Here λ_{arr} is an arrival rate in standard queueing notation, not the shadow price $\lambda_r(t)$ used elsewhere in this chapter. If we monitor actual waiting time $\hat{W}(q,r,t)$ for resource r at time t (averaged or instantaneous), and we know or estimate service rate μ_r , we can invert that to obtain ρ_r , the congestion level.

A simple heuristic is to set the shadow price $\lambda(r,t)$ proportional to wait time. For example, if the wait for a resource is three days, λ is high. To maintain dimensional consistency:

$$\lambda_r = c_r \cdot \hat{W}_{\{q,r,t\}}$$

and then scale it into the same “currency” as Φ (affect units per action). Essentially, a longer wait suggests that using that resource imposes a larger penalty in terms of foregone well-being (since waiting represents suffering or delay in balancing). For multiple servers (M/M/c), one can apply the Erlang-C formula for adjustment.

In practice, we may use a standardized congestion index when direct queue modeling is impractical; see next part B.

B. Composite congestion meter: When direct queue formulas are too strict (and real systems not Poisson, etc.), we can construct a composite index of congestion from several signals:

For each resource r , define something like:

$$\text{Congr}(t) = z(\text{wait}) + z(\text{denials}) + z(\text{utilization}) - z(\text{idle_slack})$$

This is an example formula. We take measures: average wait time (z-scored), fraction of requests denied or delayed (z-scored), utilization rate (percentage busy), and maybe idle or slack time (with a negative sign because more slack means less congestion). Sum those standardized components to get a single number $\text{Congr}(t)$. Then we set:

$$\lambda_{rt} = w_r \times \text{Congr}(t),$$

where w_r is a weight learned, for example, by fitting a hierarchical model to observed outcomes. Conceptually, λ acts as a latent variable linearly related to observable congestion indicators. This approach is useful in domains such as psychosocial or care resources, where direct queue metrics are unavailable—for instance, estimating “emotional availability” through proxies like staff-to-patient ratio or number of volunteers per client.

- Hierarchical model approach: One can fit patient- or participant-level outcomes (e.g., whether a high-draw request is made) on these z-scores, with w_s as coefficients, and partially pool across sites to obtain overall calibration. Alternatively, principal-component or factor analysis can combine congestion signals into a single latent factor.

C. Experimental manipulation of λ : We might also design experiments to see effects of pure resource scarcity:

- *Exogenous load*: Intentionally throttle access or flood a resource with dummy demand in an experiment. For instance, in a simulation scenario, randomly assign some participants to have slower response from a support channel (like “the counselor is helping others, you have to wait”) vs. others get immediate response

- thereby raising λ artificially for one group. Or hire confederates to occupy a queue in a field study setting to see if others adapt behavior (this requires ethical caution).
- *Price posting:* If ethically possible, display a live “busyness meter” or queue length to participants and see if their choices change accordingly. For example, an app could show “therapist available in 5 minutes” vs. “therapist busy, next slot 1 hour” and measure if people opt for self-help options more when wait is long. The assumption is if they respond to that info, it’s evidence they internalize λ . One must check that this effect is beyond just stress from seeing a wait – maybe measure self-reported stress.

D. Priority exception measurement: We propose a form to capture the horizon-priority effect in data: does having a short horizon reduce the effective penalty of λ for an individual? If so, we’d see something like: short-horizon individuals still use high- λ resources more than expected.

One way to model this: modify the penalty term to $(\lambda_{rt} - \kappa_r H_t^{-1})^+$, meaning subtract κ_r / H from the resource cost, not letting it go below zero. Here κ_r is a parameter that denotes how much priority boost per unit of $1/H$.

We’d estimate κ_r by looking at “exceptions”: e.g., in congestion peaks, do those with shorter horizons indeed get through more? For instance, in hospital data, during times of high ICU demand, are patients near end-of-life still being admitted at higher rates than others? A positive κ would indicate yes.

Practically, one can include an interaction term in a model: e.g., logistic regression for resource use that has terms like $\theta_\lambda \lambda + \theta_H (\Phi \cdot H^{-1}) + \theta_K (\lambda \cdot H^{-1})$.

A significant negative θ_K (since high $1/H$ means short horizon) on the penalty would correspond to a positive κ in our concept (reducing effective penalty for short H). We can fit such multi-level models and see if that term is needed.

In estimation, we might find that for certain resources, $\kappa_r > 0$, meaning indeed horizon priority occurs (short-horizon folks effectively face a smaller penalty for using it – presumably because others yield or system prioritizes). If $\kappa_r \approx 0$, no priority effect; if it’s negative (which would be weird, meaning short-horizon face even *bigger* penalty), that would falsify horizon priority.

6.4.4 Joint estimation in behavior and brain

Now, how to incorporate these into behavioral and neural analyses:

Behavioral choice model: We can formalize a choice model (for say discrete choices among options) as:

$$\Pr(u_t) = \text{softmax}(\theta_U U(u_t) + \theta_C C(u_t) + \theta_\Phi \Phi(u_t; L_t, H_t, R_t) + \theta_H \Phi(u_t) H_t^{-1} - \sum_r \theta_{\{\lambda_r\}} \lambda_{r_t} \Delta_r(u_t)).$$

This is basically a utility + conflict + compensability + horizon*compensability + resource cost model. We'd fit parameters θ to data (likely hierarchical Bayes or MLE). Key tests: is $\theta_H > 0$ significantly? (i.e., does adding the $\Phi \times 1/H$ term improve fit?), and are $\theta_{\lambda_r} > 0$? (costly signals for resources).

We might compare such models with and without those terms to see if including horizon and λ factors yields better predictions (e.g., via LOO or WAIC improvements as per our once-per-chapter mention).

Neural GLM (for fMRI or EEG regressors): For brain data (like BOLD signals in ROI or single voxels), we can include these same components as parametric modulators:

$$y_{ROI}(t) = \text{Backbone}(t) + \gamma_\Phi \Phi(u_t) + \gamma_H [\Phi(u_t) H_t^{-1}] + \sum_r \gamma_{\{\lambda_r\}} \lambda_{r_t} \Delta_r(u_t) + \varepsilon.$$

where “Backbone(t)” includes basic task regressors (like standard GLM regressors: stimulus onsets, etc., plus maybe utility and conflict if they are separate). We then ask: do we see significant γ for these terms in hypothesized regions?

For example, we predict in vmPFC: $\gamma_\Phi > 0$ (value signal for Φ) and $\gamma_H > 0$ (value signal amplified with horizon shrink). In rIFG: maybe significant $\gamma_{\{\lambda\}}$ or combined effect correlating with braking signals. ACC: possibly both horizon and λ signals.

We'd include such regressors in design matrices and see cluster activations or ROI betas.

State-space layer: Ideally, we integrate estimation of H_t (and possibly λ_t if we treat it also as latent from multiple cues) into a hierarchical model for behavior and neural data. This is advanced: e.g., one could do a joint model where H_t is a latent process as described, and the behavioral GLM uses H_t in it, and you do a big Bayesian fit to estimate both the latent trajectory and the weights simultaneously, propagating uncertainty. Or one could do a two-stage: first estimate H_t as earlier, then plug in mean estimates into behavior analysis.

In a fully Bayesian joint model, one can also incorporate prior knowledge (like we usually know horizon is decreasing, etc.) and run MCMC or variational Bayes. We'd then do posterior predictive checks to ensure the model captures key patterns and out-of-sample performance to ensure no overfit.

For complexity, we might not always fully couple them – sometimes sequential estimation is fine if uncertainty is small.

The key advice: always report how well these models predict held-out data, because it's easy to add many terms and overfit. We want to see that including H and λ states genuinely improves prediction of choices or brain signals beyond simpler models. Our pre-registration would say something like: we'll compare a backbone model vs. backbone + Φ vs. backbone + $\Phi + \Phi \times H^{-1} + \lambda$ terms, and expect an improvement in, say, log-likelihood or R^2 when we add these, if our theory holds.

6.4.5 Feature construction for Φ (with horizon/price hooks)

Recall from earlier chapters (especially 5.2.5) that $\Phi(u)$ – the feasibility-of-compensation score – can be built from component features. We need to ensure those features interact properly with horizon and congestion.

We propose constructing Φ approximately as:

$$\Phi(u_t) \approx w_1 \text{ReliefGain}(u_t) + w_2 \text{RepairGain}(u_t) - w_3 \text{Irreversibility}(u_t) + w_4 \text{OptionFlex}(u_t) - \sum_r \lambda_{rt} \Delta_r(u_t).$$

This mirrors what we saw qualitatively. The last term is explicitly subtracting resource draws – effectively integrating the penalty into “effective Φ .” But in practice, for clarity, we often keep the resource part separate in modeling (so we don't double count λ between this formula and the main equation).

Important: when horizon is short, w_3 (irreversibility's weight) and w_4 (flexibility's weight) should effectively be larger. We can achieve this by adding interaction features:

- Irreversibility $\times H^{-1}$: feature = $(\text{Irreversibility} \cdot H^{-1})$. This will capture that irreversibility is especially bad when little time remains.
- OptionFlex $\times H^{-1}$: feature = $(\text{OptionFlex} \cdot H^{-1})$. This captures that flexibility is extra valuable near the end.

We include these in the design matrix for choice models if needed (though they might be partly redundant with the overall $\Phi \times H^{-1}$ term, depending on specification – careful not to over-collinearize). Alternatively, incorporate them by making w_3, w_4 functions of H in a Bayesian model.

Similarly, under congestion, we might consider an interaction like:

- RepairGain $\times \lambda_r$: If certain repairs become more attractive substitutes when primary channels are busy, we might include an interaction (e.g., in a hospital, when formal therapy sessions (resource) are unavailable, maybe peer support (a

type of repair action) becomes more likely; that's an interaction effect). Not to complicate too much, but just acknowledging.

We will preregister all feature definitions to avoid cherry-picking. For example, define Irreversibility as a 0/1 or scale per action beforehand, OptionFlex similarly (maybe via questionnaire or rational coding), etc., and not tweak them post-hoc. The choice of features should be grounded in theory and ideally simple enough to be transparent.

We'll compute these features from available metadata: e.g., for a given action in an experiment, we know if it's reversible (maybe we set that in design), or in retrospective data, we might classify events as reversible or not. In telemetry, OptionFlex might be approximated by whether the person can later change course after doing that action (like sending a message vs. posting a public rant – one is more retractable than the other).

6.4.6 Negative controls and instruments

To avoid fooling ourselves with spurious effects:

- Sham congestion: We should verify that people respond to *real* congestion, not just any sign. So include a condition where we display a “busyness” indicator that is actually fake (like a random or baseline level) while actual availability is constant. If our measures detect an effect under real congestion but not under sham, that's good. If they react even to sham, maybe they're just reacting to the *perception* of busyness (which is still an effect, but then our λ measure should incorporate perceived, not actual, state – in any event, a clue).
- Placebo horizon: Similarly, use a non-credible deadline prime as a negative control. E.g., tell participants a deadline that they don't actually believe (maybe too absurd or we then reveal it was a joke), and see that it *does not* change behavior. If a non-credible prime does nothing but a credible one does, it assures us it's the belief of horizon that matters, not some generic announcement.
- Exclusion windows: Identify times or contexts where λ is administratively fixed to zero (like enforced protected time when no interruptions or when additional resources are guaranteed). These periods serve as natural experiments to check that when $\lambda \rightarrow 0$, we see no co-movement in menus due to resource competition. For example, if a hospital declares a daily quiet hour (no loud activities allowed), that's a period of artificially low λ for noise. We'd predict less compensatory skew outside that hour than inside, etc., or at least no congestion effect during it because we removed the variation.

The idea is to have conditions where horizon or congestion *should* theoretically have no effect (if our interpretation is correct, because either they're not present or not believed).

If we still see effects then, something's off (either an unaccounted factor or a flaw in our model).

6.4.7 Power and simulation-based calibration

Given these tests and models are complex, we have to ensure our study is sufficiently powered to detect the effects or fail patterns of interest:

We will simulate agent data under plausible parameter values (with and without horizon/price effects) to see how large a sample or how many observations are needed to reliably recover parameters like θ_H , θ_λ etc. This means generating synthetic choice sequences or outcomes with, say, $\theta_H = 0.2$ vs. 0, etc., adding noise and analyzing with our planned methods to see if we'd catch significance.

We target typically $> .9$ power to detect moderate effects (e.g., standardized coefficients in the 0.1–0.2 range) with a given N. For example, our rough calculation might say: to detect $\theta_H \sim 0.15$ with 90% power, we need ~80 participants each doing 300–600 trials in an experiment, or N ~150 patients each tracked for, say, 150 days in telemetry, etc. We'll aim for those ballpark numbers (adjust if multi-level structure reduces independent info).

Our simulation calibration will also check type I error (we want to ensure our model/analysis isn't frequently giving false positives under no-effect scenarios – if it does, adjust model complexity or use more conservative inference like hierarchical modeling to borrow strength properly).

6.4.8 Typical pitfalls (and fixes)

When implementing these analyses, watch out for:

- Confounding arousal with horizon: If a horizon prime inadvertently also increased stress, one might wrongly attribute behavior change to horizon when it's actually just stress. Fix: Always include measures of arousal (pupil, HRV) as covariates or do primes that are designed not to raise arousal (like information-based primes rather than time pressure with alarms).
- Collinearity of λ with stress: In real life, high congestion often causes stress. If we see changes under high λ , is it because of λ itself or because everyone's stressed in a crowd? Fix: We can attempt to measure stress separately (self-report or physiological) and orthogonalize λ with respect to a stress index (regress one on the other and use residual) in analysis. Or design some conditions where high load is present but individuals not stressed (maybe highly trained staff can handle load calmly) to dissociate.

- Noisy H_t meter: If our estimation of H_t is very noisy (like subjective reports all over the place, or prognostic models uncertain), then any effect of H might be diluted. Fix: Use state-space fusion to get a more stable estimate, or incorporate multiple indicators (the more signals, the better we can filter out noise). Also possibly widen the set of observations: e.g., include family's estimate of patient's time left as another perspective.
- Overfitting the backbone model: If we throw in too many nuisance terms (like a huge set of polynomial trends, etc.) we might soak up variance including the effect we care about, or if we baseline-detrend too aggressively, we might remove real signal. Fix: Preregister a parsimonious nuisance model (maybe linear effects of known confounds like time of day, fatigue, etc.) and stick to it. Demonstrate incremental value of QS terms by held-out prediction as we keep stressing – not just in-sample significance.

These precautions ensure we aren't misled by method artifacts.

6.4.9 Reporting checklist

When we publish results or share data, to be transparent and to allow others to reproduce and critique fairly, we commit to include:

- Exact definitions of H_t , λ_{rt} , and Φ features. (So others know what we meant by “short horizon” or “irreversible” precisely – e.g., H_t defined as median survival days remaining as of that date, etc., Irreversibility coded 1 for actions that permanently use resources or have lasting harm.)
- All nuisance models and comparisons: We will report the results of backbone vs. +QS model fits, including differences in LOO/WAIC or other information criteria. If we did many comparisons, we'll present those in supplement so you can see if only some analyses show improvements or all.
- Posterior intervals or confidence intervals for θ_H , θ_λ (the horizon and congestion effects) along with partial dependence plots to illustrate how predicted outcomes vary with horizon length or congestion level. E.g., a graph showing predicted probability of a repair action when $1/H$ is low vs. high, holding other things constant – with uncertainty bands.
- Instrument validity checks and negative control outcomes: E.g., we'll report what happened in sham conditions, what happened to arousal levels in horizon primes. If something funky happened (like a horizon prime also raised heart rate by 10 bpm), we'll disclose that and how we adjusted analysis.
- Site-replication summary or plan: If we did study at multiple sites or cultures, we'll either provide pooled vs. separate results or note if one site deviated. If only one

site, we will be clear that results need replication elsewhere. Possibly mention if follow-ups are underway.

This thorough reporting ensures that any claimed detection of horizon or shadow price effects is credible, and any failures to detect are interpretable (not hidden by lack of power or unreported quirks).

6.4.10 Fail pattern: specific to H_t and λ_t

Finally, in line with falsifiability, we outline the key failure modes specifically for the theory that horizon and congestion factors matter:

- Horizon null: Estimated θ_H (or analog) is indistinguishable from zero in behavior and brain models even after controlling for confounds. That is, adding $\Phi \times H^{-1}$ terms yields no improvement; no ROI shows any effect of horizon scaling; short-horizon conditions yield no difference in choices relative to long-horizon under matched utility/conflict. Also, interactions like Irreversibility $\times H^{-1}$ show nothing. This would mean shrinking horizons don't actually intensify balancing in measurable ways.
- Price null: Behavioral and neural parameters for λ are zero – people act as if they don't care about congestion. In data, no difference in choices between low-load and high-load situations (after controlling direct utility differences). fMRI: no ACC or rIFG signals correlating with λ , etc. Also, no “menu co-movement” in naturalistic data when resources are scarce (people continue as usual, no shift to substitutes). This would imply the shared-resource constraint part of QS is not operating (maybe people are either oblivious to congestion or always operate as lone agents).
- Priority null: No evidence of horizon priority in resource allocation. If short-horizon streams do *not* retain access in spikes – e.g., in ICU data, patients near EOL are triaged no differently or even less favorably during surges, or in experiments, short horizon participants drop out of resource contention just like others – then $\kappa_r \sim 0$. It suggests the system doesn't in fact preserve feasibility for those who cannot defer, at least not systematically.
- Rival sufficiency (for horizon/ λ): We find that a simpler model (like one with just utility, conflict, maybe a generic time preference) fits as well as ours. For instance, maybe a reinforcement learning model with a discount factor and risk aversion can mimic all observed behaviors without explicit H or λ terms. If so, adding H or λ doesn't improve predictions or fit, meaning those concepts aren't needed – they might be epiphenomenal or already accounted for by other known constructs.

If any two of those patterns are clearly and replicably present (especially horizon null and price null together), it would force a downgrade: we'd say maybe endgame intensification is not a universal law but a situational tendency, or we mis-specified something big. If three or more show up, we'd consider abandoning the claim that H and λ have any mechanistic role, focusing perhaps on simpler explanations (like classical decision theory or something outside QS).

(We precommit to those thresholds in our preregistration as described.)

At this point, we have laid out how to rigorously capture and test the intuitive claims from earlier in the chapter. We will use these tools in Chapter 10–13 when we actually gather and examine evidence. The goal is clear: find out if real data supports the idea that as time gets short or resources get tight, the “Law of Fairness” dynamics actually manifest – or discover where it fails.

6.4.11 Where we go next:

Definitions only matter if they survive instruments. 6.5 maps concrete measurement plans across fMRI/EEG, autonomics, time perspective, sleep/dream metrics, and naturalistic telemetry, aiming for a lean, within-subject battery that respects burden and privacy.

6.5 What to Measure (EEG/fMRI, Time Perspective)

This section serves as a measurement blueprint for detecting horizon-sensitive dynamics of the admissible set. We list concrete meters to use (from brain signals to self-reports), how to preprocess and analyze them, and what preregistered effects to look for in each. The guiding rule is: if a metric's changes can be explained by ordinary factors like utility, conflict, arousal, or habit *alone*, then it's *not* evidence for the Queue System – we only count it if adding the horizon (H) or compensability (Φ) factors gives extra predictive value. Thus, every analysis assumes a strong backbone model control, and then asks: “Do horizon-scaled compensability or congestion terms improve prediction beyond that?”

We organize by layer – brain, body, and “mindset” – reflecting different vantage points on the same process.

6.5.1 fMRI: regions, regressors, and contrasts

For functional MRI studies (or other imaging of regional brain activity):

- Core ROIs: Based on earlier theory, we focus on regions:
 - vmPFC/OFC (ventromedial/orbitofrontal cortex) – valuation.
 - ACC (anterior cingulate) – control/cost monitoring.
 - rIFG + STN (right inferior frontal gyrus and subthalamic nucleus) – braking/inhibition.
 - Anterior insula – interoceptive salience.
 - (Optional exploratory: hippocampus for memory of policies, dmPFC/TPJ for social modeling – these latter for interactions maybe).
- Task structure: We design choices or tasks in the scanner where:
 - Options are *equated on immediate utility and conflict* but differ in compensability features (i.e., differences in Φ). For instance, one choice might be a Repair option vs. an Indulgence option that both have equal immediate reward points. Or a Reversible vs. Irreversible action with same payoff. This isolates the effect of compensability (Φ).
 - We include a horizon instrument: e.g., some trials or blocks where a countdown or context implies fewer future opportunities (deadline pressure) vs. others with plenty of future. Or frame some choices as “near end of game” vs. “mid game” in instructions.

- Possibly a congestion instrument: e.g., show a cue that many others are using a resource or there's limited availability (if simulating a social scenario).
- First-level GLM design: For each trial, include parametric modulators:
 - Utility (U) – possibly as the primary regressor of interest.
 - Conflict (C) or some entropy/difficulty measure.
 - Φ (compensability score of the chosen option, or difference if two options).
 - $\Phi \times H^{-1}$ interaction modulator (value of chosen option weighted by $1/H$ at that trial, if horizon varies by trial).
 - $\sum_r \lambda_r \Delta_r$ (total resource cost) if applicable, perhaps separate modulators if we want region-specific effects.
 - Plus standard nuisance regressors: motion parameters, physiological noise, run drift terms, etc.

We preregister *ROI-level predictions*:

- vmPFC: we expect positive modulation by Φ and by $\Phi \times H^{-1}$. In other words, in vmPFC BOLD, $\gamma_\Phi > 0$ and $\gamma_H > 0$ (the horizon gain signal). This would mean vmPFC encodes the value of compensable options, especially under short horizons (amplified signal).
- rIFG/STN: we expect negative modulation by Φ (they activate more for low- Φ actions because they impose brakes) and positive modulation by $\Phi \times H^{-1}$, meaning they brake even more for low- Φ under short horizons. Alternatively described: high- Φ actions produce less braking (less IFG activation), especially if horizon is short; low- Φ actions produce strong braking signals when the horizon is short.
- ACC: expect positive modulation by λ (ACC signals conflict/cost when resource cost is high) and perhaps ramp with shorter horizon for uncompensable continuations (maybe an interaction of cost with horizon).
- Insula: perhaps positive modulation by Φ as well – reflecting the “rightness” of high- Φ decisions beyond just arousal. But insula is tricky; likely it could track internal states with horizon too.
- Model comparison: We plan to report whether including Φ and $\Phi \times H^{-1}$ modulators improves model fit (e.g., via cross-validated log-likelihood) beyond a model with just U, C, etc. This is in line with our once-per-chapter OOS mention

requirement – here it's explicitly stated to report LOO/WAIC or similar for models with vs. without those modulators.

- Multivariate checks: Besides univariate GLM, we can do Representational Similarity Analysis or decoding:
 - *RSA/decoding*: classify brain patterns for conditions like “Reparable vs. Irreparable” decisions. Prediction: classification accuracy should increase as $H \downarrow$ – meaning, as horizon shrinks, brain states for, say, choosing a compensable vs. an uncompensable option diverge more clearly, because system is strongly differentiating them.
 - *Connectivity*: test if functional connectivity between vmPFC and rIFG/STN increases for high- Φ choices near closure (implying coordination between value and braking circuits), and ACC-rIFG coupling increases for low- Φ near closure (ACC flagging cost and rIFG implementing brake).

We will preregister these multivariate and network signatures as exploratory but theoretically expected.

6.5.2 EEG/MEG: time-resolved signatures

Why use EEG/MEG? Horizon effects likely have a timing dimension: they might alter the dynamics of decision-making on the scale of hundreds of milliseconds (e.g., how quickly inhibitory signals kick in when an action is deemed bad under short horizon).

We focus on known electrophysiological signatures:

- Channels/components of interest:
 - Fronto-central theta (4–7 Hz), often from midline electrodes ~ ACC region – marker of conflict/cost monitoring.
 - Beta-band (~15–25 Hz) suppression in motor areas or rIFG region – marker of motor inhibition (the classic finding: “beta rebound” or synchronization when stopping).
 - Late Positive Potential (LPP) in parietal sites – often related to sustained attention/valuation of emotional stimuli, might indicate processing of meaningful cues.
 - Contingent Negative Variation (CNV) or CPP (central positivity) – related to evidence accumulation especially under deadlines.
- Contrasts to examine:

- Φ effect on stopping: E.g., in a Go/NoGo or stop-signal task where “Go if action is compensable, NoGo if not” scenario, we predict:
 - Reduced beta desynchronization (i.e., less motor preparation) and shorter stop-signal reaction times (SSRT) for high- Φ “Go” decisions (ease in going), and conversely enhanced beta and longer SSRT for low- Φ “NoGo” especially as H shortens. Put simply, stopping an imprudent action should become easier (stronger inhibition) as horizon shrinks.
 - Horizon interaction in time-frequency: Theta power slope vs. time-to-deadline might increase for low- Φ sequences (like ACC shows more rising conflict signal as deadline nears), and LPP amplitude might be larger for closure-related decisions near the end (indicating heightened salience processing).
- Preprocessing and stats: We will pre-register:
 - Filter bands (theta, beta).
 - ICA for artifact removal, define epochs (like decision onset to outcome).
 - Use cluster-based permutation tests or hierarchical Bayesian models for time-frequency differences, given multiple comparisons.
 - Possibly plan blocked cross-validation across sessions to avoid overfitting noise patterns.

Expected outcomes:

- Clear differences in brain dynamics: e.g., at a fixed time before a deadline, low- Φ tasks show a steep increase in frontal theta that high- Φ tasks don’t (meaning more conflict under horizon for bad tasks).
- Perhaps an LPP more pronounced when making a reconciliation decision under short horizon (sign of deeper processing).

6.5.3 Autonomic and interoceptive meters

The claim “the admissible set is felt” means we expect bodily signals to reflect ease vs. heaviness beyond generic arousal.

We measure:

- Pupil diameter (tonic and phasic).
- Skin Conductance Level/Response (SCL/SCR).

- Heart rate (HR) and heart rate variability (HRV, especially high-frequency component reflecting parasympathetic activity).
- Respiration rate and variability.
- Maybe skin temperature if relevant (stress can cool extremities).
- Micro-EMA: participants can give a 1–2 word descriptor of how a decision felt (“heavy”, “right”, “stuck”, “light”, etc.) after key choices – capturing phenomenology briefly.

Predictions:

- Before choice: for a high- Φ option near closure, we expect lower tonic SCL and higher HF-HRV (indicative of calm/ease) relative to baseline; for a low- Φ at short horizon, higher sympathetic signs (SCL up) and lower HF-HRV (anxiety). In other words, right moves feel safe and come with a parasympathetic “OK”, wrong moves feel pressing and jarring.
- After commitment: physiologically, high- Φ choices should show faster recovery (autonomic settling) especially as $H \downarrow$. E.g., after deciding on a closure action, heart rate might quickly normalize, whereas after indulging when time is short, maybe HR remains elevated (sign of lingering internal conflict).
- These effects should remain even after controlling for effort or expected value differences (so it’s not just “this decision was easier or more rewarding”).

We’ll measure these around decision points. Possibly use wearable sensors if doing field study (like chest strap for HRV, EDA patch for SCL, pupillometry in lab tasks).

We must ensure to isolate these from general arousal: controlling for physical activity, etc.

6.5.4 Bioelectric Field Coherence (System-Level Regulation)

Beyond localized neural activation and autonomic markers, there is a deeper physiological layer that may participate in constraint dynamics: endogenous bioelectric fields.

All living tissues maintain voltage gradients. Neurons communicate via action potentials, but bioelectric regulation is not limited to neurons. Epithelial layers, organ systems, and developing embryos rely on patterned voltage gradients and ionic fluxes to coordinate growth, repair, polarity, and structural stability. These fields function as system-level constraint architectures — maintaining coherent form and suppressing runaway divergence.

This matters for horizon scaling because conscious experience emerges from an electrically embedded organ operating within body-wide electrochemical regulation. The brain does not function as an isolated circuit board; it is immersed in and influenced by distributed electrical gradients across cortex and body.

The Law of Fairness does not claim that bioelectric fields “store” ledgers or encode moral information. It requires only that biological systems possess layered constraint mechanisms capable of resisting indefinite experiential divergence. Bioelectric regulation provides a plausible substrate for large-scale stabilization beyond synaptic learning alone.

If experiential imbalance corresponds to prolonged dysregulation in large-scale neural coordination — for example, excessive variance, runaway excitation, or unstable integration — then global electrical coherence should change as horizons shrink. Specifically, as H_t decreases and admissible sets narrow, we predict measurable shifts in large-scale electrophysiological stability.

Measurable predictions:

- Increased global coherence (e.g., phase-locking value, coherence metrics, or cross-frequency coupling stability) as closure approaches in well-supported individuals.
- Reduced broadband variance in EEG/MEG power spectra during final convergence windows (variance compression at the electrophysiological level).
- Stronger coupling between frontal control networks and large-scale oscillatory coherence during short-horizon high- Φ decisions.
- In end-of-life contexts (with intact comfort channels), increased slow-wave synchrony or stabilized alpha/theta dynamics relative to mid-life baseline, independent of sedation effects.

These predictions are correlational signatures. They do not imply teleology, cosmic intention, or mystical “energy fields.” They are testable physiological correlates of a system converging toward reduced variance under finite-horizon constraints.

Failure condition: If shrinking horizons show no systematic change in large-scale electrophysiological coherence beyond what is fully explained by stress, medication, fatigue, or standard predictive coding accounts, then bioelectric regulation does not add explanatory value to the constraint model.

This section therefore expands the measurement program. If the Law of Fairness is real, constraint should be detectable not only in decision weights and regional activation but in system-level electrical organization.

6.5.5 Time perspective and horizon meters

We partially covered measuring H_t in 6.4.2, but here focusing on self-report scales:

- EMA items (preregistered): For example:
 - “For this decision, it feels like I have: [hours / days / weeks / months / more than months] left to make things right.”
 - “This option would be easy to reverse later: [Likert 1–5].”
 - “If I didn’t take action now, I would lose routes to finish: [Likert].”. These directly gauge perceived horizon and the person’s sense of urgency or flexibility at that moment. We’ll ask them right after or during tasks in micro-surveys.
- Scales: Use validated short forms like:
 - Future Time Perspective scale.
 - Consideration of Future Consequences (CFC). We will not treat these as substitutes for actual horizon but as trait covariates. For example, a highly future-oriented person might behave as if horizon is longer (or might prepare better, affecting results).
- Psychometric target: We can try to build a latent H_t factor from multiple questions
 - + cues, as mentioned. Also, we’ll check that horizon effects we find are not only coming from a subset of people (e.g., only those who are generally future-minded)
 - so we’ll test that $\Phi \times H^{-1}$ effects hold *within each bin* of trait time perspective, not just between people. Essentially ensure it’s a dynamic effect, not just a static trait effect mislabeled.

6.5.6 Sleep and dream metrics (night shift of horizon)

A fascinating “night shift” aspect: if the mind tries to balance ledger overnight, we should see horizon influences in sleep architecture and dream content.

We measure:

- Polysomnography or sleep tracker data focusing on:

- REM density (how much rapid-eye-movement per unit time, or theta power during REM – REM associated with emotional processing).
 - SWS (slow-wave sleep) power (deep sleep, restorative).
 - Number of arousals or awakenings (fragmentation indicates stress).
- Dream-affect integral D (some measure of the emotional tone or content of dreams), via:
 - Brief awakenings method: wake the person after REM periods to get a report (only in research-friendly settings).
 - Morning dream logs with rating of how positive/negative the main emotion was.
 - Or text analysis of dream reports (if detailed) for reconciliation themes, etc.

Predictions:

- Day-night inversion: If horizon is short, we predict that dreams will counterbalance daytime drift. E.g., after a day with negative ledger change, nighttime dreams produce positive imagery (and vice versa) – especially stronger when H is small (because more urgent counterweight needed). Mathematically, maybe $D \approx -\alpha(H) \Delta L_{\text{day}}$ with $\alpha(H)$ increasing as $H \downarrow$. That is, the dream affect D is inversely proportional to the daytime ledger change, with a factor that grows when horizon is short (meaning the shorter the horizon, the more dreams do the opposite of daytime emotion).
- Next-day tilt: A night with certain dream patterns (like high REM after a bad day) should yield a measurable improvement in morning outlook or choices: “morning advantage” for high- Φ options after such nights. E.g., if someone had a tough day (negative affect) but then a REM-rich night with maybe intense dreams, next morning they might more readily choose a compensatory action than if they didn’t have that dream rebound.

We plan to correlate dream reports with next-day behavior: do reconciliation dreams correlate with making reconciling actions the next day? Already some anecdotal evidence, but we can measure frequencies.

We must control for baseline mood, medication (some meds suppress REM), etc.

6.5.7 Naturalistic telemetry: phone, text, mobility

Beyond lab measures, we can use smartphones and wearables to observe behavior in real life to detect horizon effects:

- Signals to collect:
 - Repertoire index: count of unique action types per day (calls made, distinct contacts spoken to, distinct locations visited, app categories used, etc.).
 - Repair proportion: fraction of communications or activities directed to “repair” social ties or comfort. E.g., sentiment analysis of text messages for positive sentiment to close contacts, or categorize call recipients (family vs. work). Or track how many supportive vs. adversarial exchanges.
 - Stickiness: measure completion vs. abandonment rates of tasks: e.g., how many drafted messages were sent vs. deleted (if we can get that), how many phone calls started vs. hung up quickly, how many to-dos deferred vs. done.
 - Congestion meter: things like how often did a person experience waiting (perhaps from phone usage patterns like waiting on hold, or attempted but failed resource uses like tried scheduling an appointment and got none – if such logs exist). Or for health, maybe how often they press call nurse button and how long until resolved if we have that data.
- Predictions (which echo earlier ones, but now in daily life):
 - As objective or perceived horizon shortens (like as someone nears retirement, or an older person in final year vs. earlier, or a student near finals vs. start of term), we should see Repertoire narrowing and repair proportion rising. For a patient in hospice, in their last week they call basically only close family (few contacts) and mostly positive/supportive content; a month before, they might have been engaging in more variety.
 - Stickiness asymmetry: Should grow near closure. We can quantify maybe by “persist/abort ratio” for various tasks. Expect: tasks classed as high- Φ have higher completion rates (persist once started) than low- Φ tasks, especially among those near end or on days close to significant endings.
 - Menu co-movement across people: When λ is high (like a community event overload or hospital crowded), do many people’s behaviors shift together towards substitutive actions? We can look at correlation of individuals’

activity changes. Also, horizon-priority exceptions: short-horizon individuals' patterns should deviate (they still do some high-draw acts others forego).

All these can be tracked with appropriate privacy and consent. We emphasize privacy-preserving analysis (like embedding sentiments rather than reading content, etc.).

6.5.8 Minimal lab battery (2 hours, within-subject)

We propose a concise lab protocol to capture key phenomena within a single session per participant:

1. Horizon-manipulated choice task in fMRI (or EEG): e.g., a decision game with blocks where sometimes a countdown is shown (short horizon block) vs. not (long horizon block), utilities matched. Meanwhile record pupil as well. (Say 40 minutes.)
2. Stop-signal / Go-NoGo task with tags for reversibility and repair: embed signals like “Go if reversible action, NoGo if irreversible” to measure inhibitory control differences. Collect EEG beta/theta, measure SSRT differences by condition (maybe 20 minutes).
3. Autonomic block: e.g., have them watch or imagine scenarios: one of taking a relief action vs. taking an indulgence, with and without time pressure. Record HRV, SCL continuously (maybe 15 minutes scenario imaginations with rest periods).
4. EMA + time perspective surveys throughout: pepper short questions between tasks about how time feels, then at end do trait surveys plus an “evening reflection” for drift (simulate day’s end). (15 minutes across session.)
5. (If possible) Sleep/dream measurement: Unlikely in a 2-hr lab, but maybe ask about last night’s sleep/dream as baseline, or offer an optional night extension.

This battery's pass criteria (what we hope to see, else it's a fail):

- vmPFC shows $\gamma_\Phi > 0$.
- rIFG/ACC show a $\Phi \times H^{-1}$ interaction in EEG/fMRI.
- Autonomic: show an effect of action type (ease vs. heavy) dissociable from effort (e.g., HRV higher for repair vs. indulgence controlling for physical effort).
- Behaviorally: in the tasks, see that sequences that are low- Φ stall more as horizon shrinks (like more NoGo success for low- Φ near deadline).
- And crucially, out-of-sample prediction improves when including H and Φ terms across these tasks.

We only consider the law supported if *all* primary outcomes show pattern in predicted direction with out-of-sample validation. That's a high bar, but appropriate.

6.5.9 Analysis principles (to avoid fooling ourselves)

We will follow strict principles (some repeated from 6.4 but worth consolidating):

- Backbone-first: Always fit a lean model with usual suspects (utility, conflict, risk, habit, fatigue, learning effects) *before* adding any QS terms (Φ , H^{-1} , λ). See if adding them improves things – if not, we don't claim detection.
- Blocked CV: We will hold out whole days or whole subjects as test sets, not just random trials, to ensure generalization. No training on the same subject's data to test horizon effects – that could overfit idiosyncrasies.
- Negative controls: Use sham horizon and sham congestion conditions as described, expecting QS terms to have no effect there. If they still appear, reevaluate (maybe we captured a spurious effect). We won't count results as confirmatory if they also show up where they shouldn't.
- Instrument checks: Confirm through manipulation checks that when we attempted to change horizon or λ , we indeed did (subjective or objective). If not, then a null result isn't evidence against QS, it's evidence our manipulation failed.
- Multiplicity control: We define exactly which metrics are primary vs. exploratory. Primary ones get hypothesis tests (with correction if multiple), the rest are descriptive. We avoid cherry-picking peaks or timepoints – use cluster stats or whole-curve criteria. Predefine ROI's of interest for brain to avoid full brain fishing (beyond perhaps one confirmatory cluster test for each predicted region or network).
- Share code and priors: We'll collaborate openly (adversarial collaboration idea from earlier chapters) – e.g., sharing analysis code with a skeptic team for them to verify or even run their own models on our data. We might let a skeptic define a nuisance model to ensure we're not biasing something. And give them say in writing a "why this could be wrong" part if published.

The goal is to avoid any subtle bias or wishful thinking affecting results, given the extraordinary nature of the claim (we want to be extra sure).

6.5.10 Reporting template (what goes in the paper)

Finally, how we'll present results to be clear:

- Tasks, instruments, feature definitions: A table or section listing each experiment or data source, what horizon/ λ manipulation was used, how Φ features were defined, etc. (So readers know exactly what we did.)
- Backbone vs. +QS model comparison: Likely a table of $\Delta\text{elpd}_{\text{loo}}$ (PSIS-LOO) or ΔWAIC for adding Φ and adding $\Phi \times H^{-1}$ and adding λ terms sequentially, plus maybe ROC or log-loss for behavioral predictions. Also for neural: maybe information criteria or cross-validated R^2 improvement for including QS regressors.
- ROI coefficients: A figure with a forest plot of credible intervals for key γ terms (e.g., γ_{Φ} in vmPFC, γ_H in vmPFC, γ_{λ} in ACC, etc.), indicating which are significantly greater than 0 or less than 0.
- Time-frequency maps with cluster stats: For EEG, maybe an image showing beta power difference wave with significant clusters shaded.
- Partial dependence plots: Graphs showing predicted probability of a high- Φ choice as function of horizon (with and without QS terms). Or predicted ACC activation vs. congestion at different horizons, etc., to interpret interactions.
- Replication across sites or sessions: If we have multiple, maybe a meta-analytic summary, or at least note consistency. If any null, we mention with power analysis.
- Nulls with power reported: If we didn't find something, say we expected X and got null, we'll state what effect size we could have detected with our power. (E.g., "we found no effect of horizon on stickiness; sensitivity analysis suggests we'd detect a $\geq 10\%$ difference with 90% power, so any effect might be smaller or non-existent.")

Finally, a Takeaway summary in the paper: listing pattern recurrence across all measures:

- e.g., "Valuation gain for compensable options intensified as horizons shrank (vmPFC $\beta=...$, behavioral $\Delta\text{elpd}_{\text{loo}}$ PSIS-LOO=..., etc.)",
- "Braking scaled with horizon on low- Φ actions (rlFG SSRT +50ms as predicted)",
- "Somatic markers (HRV +, SCL -) indicated ease for short-horizon right moves vs. heaviness for wrong ones independent of arousal",
- "Dream analysis showed night counterweights consistent with horizon effects",
- "Many of these signals improved prediction beyond classical models"

That will make it clear: if the Law's "guardrails" exist, *the same pattern should appear in all these metrics*. If we cannot measure them clearly—after controlling for other factors—then we should question whether we have a law-level mechanism at all.

In short, this chapter set the stage for empirical tests. Next, in Chapters 7–9, we delve into constructing the Hedonic Composite Index and ensuring measurements are comparable across people—tools we'll need to fairly test these predictions in Chapters 10–13, where the evidence awaits.

6.5.11 Where we go next:

Measurements must fail cleanly when they're wrong. 6.6 enumerates null patterns and downgrade rules—the disciplined ways this program could lose—so that a negative result is just as informative as a positive one.

6.6 Fail Patterns for Horizon Scaling

Not all “null” results are equal. This section catalogs specific failure patterns that would undermine — or outright falsify — the claim that admissible choices (the “menus” of actions and thoughts) tighten with shrinking time horizons. Each pattern describes (i) what it would look like in data (Signature), (ii) plausible benign explanations to rule out first, and (iii) when the pattern would count as a strike against the theory.

6.6.1 The flat-slope null (no $\Phi \times H^{-1}$ effect anywhere)

Signature: Across tasks carefully constructed to equate immediate utility and conflict, the interaction term $\Phi \times H^{-1}$ (compensability \times inverse horizon) is indistinguishable from zero. This null effect appears in behavior (choices and reaction times) and in all key brain regions (vmPFC, ACC, rIFG/STN, insula) after including the usual “backbone” controls. Out-of-sample model fits show no improvement from adding any horizon-dependent term.

Benign explanations:

- Participants didn’t truly perceive the horizon manipulation as real or credible (e.g. they didn’t believe the “deadlines” in the task, so the horizon factor never actually changed in their minds).
- The horizon meter is too noisy – the measurement or state-space fusion of objective and perceived time-left is imprecise, obscuring any real effect.
- An arousal confound – the “deadline” conditions inadvertently increased stress or arousal, and those effects (rather than horizon per se) masked or swamped any horizon-based pattern. *Counts against if:* All the above checks are passed (participants do feel the time pressure, arousal is held constant, and horizon length is measured precisely), yet results remain null across labs. Especially damning would be if this flat result replicates in naturalistic settings where horizons truly shrink (for example, observing people near real endgames like final exam week or hospital discharge and still seeing zero change in behavior or brain measures). Such a flat-slope null, verified in rigorous experiments *and* in real-life end-of-life contexts, would directly challenge the idea that shrinking horizons intensify balancing behavior.

6.6.2 The value-only mirage (vmPFC $y_\Phi > 0$ but no control scaling)

Signature: The vmPFC (valuation center in the brain) registers extra “value” for high- Φ options (i.e. options that would significantly relieve or repair the ledger imbalance), indicating some recognition of compensability. However, control regions like rIFG/STN and ACC show no horizon-dependent “braking” on low- Φ (low-compensability) actions.

In other words, the agent's impulse control doesn't ramp up when the horizon is short. For instance, stop-signal reaction times and neural inhibitory signals (e.g. beta-band power or STN activity) fail to lengthen as $H \downarrow$ (as time runs out).

Benign explanations:

- The tasks may have been under-powered to elicit inhibitory control – e.g. too few “No-Go” or stop trials to detect a change in braking behavior.
- Inhibition signals smeared by analysis – for example, the EEG or fMRI measures of control (like beta-band or a small STN region) might be too coarse or filtered, hiding the effect.
- *Counts against if:* With adequate power and targeted inhibition tests (plenty of opportunities to brake, and fine-tuned measurements of control signals), we still see no increase in control as horizons shrink. This would imply the agent appreciates compensability in theory (values high- Φ outcomes) but isn't actually constrained by the shrinking horizon — a mere “taste” for balance without any guardrail to stop non-compensable actions. Such a result suggests the Law of Fairness isn't operating as a constraint at all in endgames.

6.6.3 The arousal substitution (effects disappear after arousal control).

Signature: At first glance, the data might show the expected pattern (e.g. stronger repair-oriented actions as deadlines loom, consistent with horizon effects). However, once you add controls for arousal or stress (e.g. include pupil size, skin conductance (SCL), or heart-rate variability (HRV) as regressors), the previously significant $\Phi \times H^{-1}$ effects vanish. In other words, what looked like a horizon-driven behavior could be fully explained by arousal or anxiety surges as time runs out. *Benign explanations:* The true horizon effect might partially operate through interoceptive arousal – meaning part of why people act to rebalance near the end is that they feel more stressed or emotionally heightened, which is actually a legitimate piece of the mechanism (not a confound).

Counts against if: Experiments are designed specifically to hold arousal constant while manipulating horizon (for example, using careful instructions or pharmacological means to keep stress levels equal), yet no horizon effect remains. In that case, it suggests that all the “end-of-horizon” behavioral changes were really just due to generic stress or fatigue factors. Horizon adds no unique explanatory power once those standard factors are accounted for, undercutting the uniqueness of the Law's predictions.

6.6.4 The congestion-only world (no horizon effect, only λ).

Signature: In this scenario, we do see some balancing behavior, but it tracks momentary congestion (λ) rather than horizon. For instance, when people’s “resource congestion” is high (lots of pending pains to compensate, represented by higher λ_t in the model), their choices shift — but an actually short remaining horizon (imminent ending) gives no extra push. All observed effects can be explained by the backlog of needed compensation, with no special endgame tightening. In data or models, we find that including terms for immediate pain/pleasure congestion improves predictions, whereas adding horizon terms does not improve anything. Short-horizon individuals get no special priority; their behavior looks no different from others once you account for how much imbalance they have.

Benign explanations:

- The studied population might lack real horizon variation – for example, perhaps almost everyone in the study is young (far from end-of-life) or, conversely, nearly all are near end-of-life, so horizon effects can’t manifest because there’s little contrast.
- Administrative or policy factors forbid horizon-based adjustments – for instance, even if someone is near the end, maybe protocols prevent any special treatment, effectively blocking the mechanism (no “triage” or priority changes allowed). *Counts against if:* There are clear differences in horizons among subjects (some truly short-term, some long-term), and context allows horizon-based prioritization, yet we still observe zero effect of horizon beyond congestion. That outcome would point to a resource-rational explanation of behavior (people just respond to the size of the pain backlog, not time left) with no genuine endgame effect. It means the Law of Fairness’ supposed “horizon constraint” might not be a necessary piece of the puzzle.

6.6.5 The non-selective narrowing (generic shutdown)

Signature: As the horizon shrinks, people do seem to slow down or narrow their activity, but not in a focused way — they simply do less of everything. Overall activity and variance drop near “deadlines,” yet crucially, they do not favor reparative actions over indulgent ones. Both low- Φ (indulgent or non-compensatory) and high- Φ (pain-relieving, compensatory) options lose their appeal equally; there’s a general withdrawal or shutdown rather than a targeted push for balance.

Benign explanations:

- Burnout or fatigue – perhaps as time runs out (or in end-of-life scenarios), people are physically and mentally exhausted, causing a broad drop in all activity that isn't specifically about balance.
- Task artifacts – maybe the experimental incentives or setup accidentally discouraged any action near the end (e.g., a penalty for acting late), so people learned to just quit trying, uniformly across option types.

Counts against if: The fatigue factor is controlled (ensuring participants are rested, or statistically accounting for wear-and-tear) and tasks are designed so that there's no disincentive to act near the end, yet we still see only a non-selective slowdown. This pattern would indicate that what looks like a horizon effect is nothing more than exhaustion or disengagement. It fails to support the idea of a discriminating mechanism that specifically preserves compensability (fairness) — instead it looks like a general collapse of effort.

6.6.6 The night-shift null (no day–night inversion)

Signature: A core prediction of the theory is that dreams act as “low-cost counterweights” — if the day’s ledger drifts negative, the night should compensate (via emotionally intense dreams, etc.), effectively inverting the trend. The Night-Shift Null means we find no such inversion: the integral of dream affect (let’s call it D) does not consistently oppose the previous day’s drift ΔL_{day} . Also, any parameter like $a(H)$ (which would measure how strongly dreams compensate as a function of horizon) shows no dependence on horizon at all. Experimental manipulations like suppressing or increasing REM sleep have no effect on next-day choices or “admissible set” tilt. In short, dreams don’t do the predicted work of balancing late-stage ledgers.

Benign explanations:

- Weak dream sampling – maybe the null result is because the way we collect dream data (self-reports, short lab stays) isn’t capturing the full picture; the compensatory dreams might be there but we didn’t detect them.
- Physiological or medication confounds – e.g., participants might be on medications that alter REM sleep, or perhaps REM signals are too subtle, meaning the experiment wasn’t truly sensitive to the dream mechanism.

Counts against if: Using strong polysomnography data and adequate sampling (i.e. really capturing dream content and affect), and controlling for meds or other sleep distortions, we still see nothing – no inversion of daytime effects, no horizon-sensitive dreaming. This outcome would undermine the “night workshop” idea central to endgame compensation. It suggests that dreams are not acting as

the hypothesized safety valve for fairness toward the end, challenging a key piece of the Law of Fairness framework.

6.6.7 The stall-free spiral (sequences don't stall)

Signature: Consider multi-step harmful sequences (like a revenge spree or an indulgence binge) that, if unchecked, lead to large ledger imbalances. The theory predicts that as the time horizon shrinks, such sequences should “stall out” – the system should throw up roadblocks (via fatigue, doubt, interruption) to prevent an irrecoverable plunge. Stall-Free Spiral means we observe no such stalling: people with short horizons continue these damaging sequences just as readily as those with long horizons. In lab tasks, long chains of self-harming or other-harming actions show no greater tendency to stop or self-correct at the last moment. Neurologically, we see no rising activation in control areas (rIFG/ACC) that would signal an increasing cost to continue the sequence as the end nears.

Benign explanations:

- Perhaps the sequences studied are too short or too unclear – maybe truly pernicious spirals weren’t captured, or it wasn’t obvious in the data what counted as a “sequence,” so the effect wouldn’t manifest.
- *Counts against if:* In experiments with clearly identified, extended sequences, and with credible horizon manipulations, we still find no difference in how those sequences play out at endgame versus mid-game. If harmful spirals don’t naturally stall more often when time is almost up, it indicates an absence of the predicted guardrails right when they should be tightest. The system isn’t stepping in to prevent last-minute irreparable damage, contradicting the Law’s expectation of protective constraints as the end approaches.

6.6.8 The cross-site inconsistency (fragile or idiosyncratic effects)

Signature: Here, any supposed horizon effects turn out to be fragile. Perhaps one lab or one dataset finds a big effect of shrinking horizons, but another lab doesn’t; or the effect only appears under very specific conditions and vanishes with minor changes. For instance, tweak the task instructions slightly, or change the analysis method, and the “significant” result either disappears or even reverses sign. In short, horizon effects, if present, are not robust across contexts.

Benign explanations:

- Local quirks – maybe the positive findings came from an unusual sample or a fluke in the analysis pipeline (e.g., a bug or an undisclosed flexibility in methods).

Counts against if: When researchers proactively address these concerns — e.g., doing adversarial collaborations (bringing in skeptics to double-check methods), using preregistered analysis pipelines, and repeating studies across multiple sites — the horizon effect still fails to stabilize. If after all that, the results are still all over the place, it suggests the phenomenon might not be real. A true law-like effect should be reliable and reproducible; if horizon effects are too idiosyncratic or fickle, it weakens the case that a fundamental “Law of Fairness” is at work.

6.6.9 The rival sufficiency (leaner models win)

Signature: A rival theory (one not involving any special fairness constraint) manages to explain all the data just as well. For example, perhaps a combination of predictive coding (uncertainty minimization) plus standard risk aversion plus fatigue can produce repertoire narrowing, some menu co-movement, even certain sleep effects that we observe, all without invoking Φ or horizon terms. Or a resource-rational reinforcement learning model with a simple cost for having too many unaddressed issues (“queue costs”) replicates the behavior. Crucially, these rival models use fewer parameters or assumptions than the full LoF/QS model. In statistical terms, they achieve better model fitness (e.g., lower log-loss, better WAIC/LOO scores) by virtue of being simpler and not overfitting. In essence, everything the Law of Fairness explains, a leaner theory can also explain – making the special “horizon” mechanism unnecessary.

Benign explanations:

- Perhaps our implementation of the Law’s model was suboptimal – e.g., the way we quantified Φ (the compensability features) was flawed, so we gave an unfair advantage to the rival model. *Counts against if:* After trying multiple reasonable ways to specify Φ and other model details, the rival still matches or beats the LoF-based model in predictive power and generalization. If adding the whole Queue System/horizon apparatus yields no clear improvement, Occam’s razor favors the simpler explanation. This would mean the “horizon scaling” idea might not be a fundamental law, but just one possible description that isn’t actually needed. In such a case, we’d have to consider downgrading the Law of Fairness from a hard constraint to maybe just a coincidental tendency — or discard it altogether if simplicity and data keep favoring the rival.

6.6.10 Telemetry nulls in real endgames

Signature: One of the strongest tests is to look at real-life end-of-life or end-of-milestone situations (not just lab tasks). Suppose we track people continuously (telemetry from wearables, smartphone data, etc.) during “final stretches” of various kinds: the last few

weeks of a semester, the weeks before a major surgery, or actual hospice care in the final months of life. The theory predicts noticeable changes (seeking closure, shifts in behavior), but a Telemetry Null means we see none of that. In these genuine short-horizon periods, the data would show no narrowing of activities, no tilt toward making amends or contacting loved ones, no change in how sticky or salient certain options are. This is despite all channels being “open” (people are cognitively intact, not heavily medicated or anything that would dampen behavior) and with factors like pain management optimized so as not to obscure behavior. Essentially, people nearing known end-points behave indistinguishably from those not near an end, in terms of seeking balance.

Benign explanations:

- Data too coarse or filtered – it could be that our telemetry isn’t fine-grained enough (missing subtle social interactions or internal states), or privacy protections stripped out the very signals that would show compensatory moves (for instance, we can’t see the content of conversations or personal reflections).
Counts against if: Using high-quality, consensual telemetry (rich data, with participants’ permission to analyze meaningful signals) across many people, we still find no differences in end-of-horizon behavior. And if this holds in repeated cohorts (multiple classes of students, multiple hospitals, etc.), it’s a serious blow. It would imply the Law of Fairness has no detectable impact even when it really should (when time is running out in reality). That would call into question the ecological validity of the whole theory.

6.6.11 Decision rule for downgrade

We have pre-committed a falsification rule for horizon scaling. Specifically: if any two of the above fail patterns replicate with adequate statistical power, rigorous controls, and in independent labs, then we will downgrade the horizon-scaling claim from a “guaranteed constraint” of the system to merely a proposed tendency. “Adequate statistical power” will be defined by preregistered minimum detectable effect sizes and cross-validated model comparison criteria, not by post-hoc significance alone. And if three or more of these patterns replicate reliably, we will abandon the horizon-scaling hypothesis entirely as a component of the Queue System. In other words, the burden of proof is high, and we won’t cling to the theory if multiple lines of evidence show it doesn’t hold up. This rule ensures we remain honest: the Law of Fairness (and its horizon effects) must earn its status through empirical success, or we are prepared to scale back our claims.

6.6.12 Troubleshooting flow (for investigators)

Even if some horizon effects are hard to detect, it's possible the fault lies in implementation. For researchers trying to demonstrate (or refute) horizon-based balancing, here is a checklist of troubleshooting steps:

1. Manipulation Check: First, verify that the experimental manipulation of horizon actually worked *without introducing confounds*. Did participants truly sense that their time or opportunities were limited (shrinking H), without merely getting more aroused or anxious? If people don't believe or perceive the horizon, or if horizon changes are always tangled up with stress, the test is invalid. *Fix:* use more credible countdowns or framing, and ensure any stress induction is equalized between conditions.
2. Meter Fusion: Assess whether your measurement of the horizon H is precise and stable. Are you combining objective and subjective indicators of remaining time in a state-space model to get a clean H_t ? If H is noisy or poorly estimated, a real effect can vanish in the noise. *Fix:* incorporate clearer objective cues (like known event timings) or improve the state-space modeling so that an individual's perceived horizon is tracked as accurately as possible.
3. Feature Audit: Double-check the definition of Φ (and related compensability features). Are the **features of “compensability” well-specified and preregistered? If Φ is too complex or post-hoc, one might miss effects or see illusory ones. *Fix:* simplify to core components – e.g. ReliefGain, RepairGain, Irreversibility, OptionFlexibility – and commit to those in advance, so analyses don't cherry-pick features that produce a desired outcome.
4. Power and Cross-Validation: Make sure the study has enough power (sufficient sample size, enough critical trials) to detect the subtle horizon effects. Use robust validation: e.g. blocked cross-validation (check generalization across different time blocks or groups of participants) and multi-site replication. If a horizon effect only appears under very specific conditions or analytic choices, it's not convincing. Ensuring it generalizes across held-out data and different environments is key.
5. Rival Model “Bake-Off”: Always pit the Law of Fairness model against a lean rival model in analysis. Include a simpler explanation (e.g., a model without horizon terms, or with just basic factors like utility, fatigue, etc.) and see which predicts data better. If a simpler model accounts for results, then the fancy horizon-based model hasn't proven itself. This forces us to show that adding the LoF/QS mechanism truly yields new explanatory power beyond existing theories.

Bottom line: Horizon scaling (the intensification of balancing behavior as time runs out) must earn its keep. We will only consider it a genuine law-like constraint if it produces clear, distinctive, horizon-contingent signatures — in decision-making, neural control signals, dream patterns, etc. — and those effects survive rigorous stress-tests (arousal controls, multi-site replication, and head-to-head comparisons with simpler models). The various fail-patterns listed above make it deliberately easy for the theory to fail under honest testing; that is by design. We do not want to embrace “horizon effects” unless they can pass these tough tests. If they consistently don’t, then the Law of Fairness needs to be revised or even rejected in light of the evidence.

6.6.13 Where we go next:

If endgame pressures are real, they should scale beyond any single person. 6.7 asks how a population-level shadow price might emerge, and how policy windows could reduce suffering without moralizing—guardrails, not steering.

6.7 Population Shadow Price and Policy Windows

Thus far we've considered horizon effects on an individual level. But what about groups or whole societies facing looming "horizons"? We can define a population-level shadow price Λ_t as the marginal urgency of preserving compensability (fair balance) across *many* streams at once. Formally, one could imagine:

$$\Lambda_t \equiv \partial/\partial R \Pr(\bigcap_i L_i(T_i) \in [-K, K]),$$

where R is some controllable resource we can allocate (for example, number of ICU beds, supply of a lifesaving drug like naloxone, emergency housing vouchers, etc.), and the probability term is the chance that *all* individuals i reach the end of their horizon T_i with their ledgers L_i within the acceptable bounds $[-K, K]$ (i.e. everyone stays "in balance" within some tolerance K). If R is discrete (beds, vouchers), interpret $\partial/\partial R$ as a finite-difference marginal gain from adding one unit of the resource at time t under a specified allocation policy. This expression is heuristic, summarizing a marginal effect of resource allocation on joint neutral-closure probability; it is not assumed differentiable in real populations and will be operationalized via discrete policy changes. Intuitively, Λ_t measures how much adding a bit of resource at time t improves the odds that everyone gets to neutral by their end.

As horizons shorten for a large subgroup of the population (for instance, an aging cohort reaching end-of-life, or a widespread crisis that could abruptly end many lives or life plans), Λ_t rises. In plainer terms, the system's "guardrails" tighten on a society-wide level. We often see this in practice: when many people are in danger of irrevocable loss, society adjusts by making certain choices more admissible and others less admissible. For example, authorities might implement temporary rules, emergency funds, or safety nets — interventions that wouldn't be justified in normal times — to help preserve balance for as many people as possible. Conversely, high-variance gambles or risky endeavors that could further threaten compensability become discouraged or outright forbidden (because there's no time to recover from bad outcomes).

Policy windows are those critical moments when a small change in rules or resources can produce a large improvement in overall compensability for the population. In a sense, they are high- Λ_t moments: making a policy move during that window (say, deploying a relief program or changing a law) can dramatically increase the fraction of lives that stay balanced. The Queue System theory predicts we should detect collective shifts around such windows. For example, we might see spikes in public attention (e.g., search trends or social media focus suddenly revolving around an issue), changes in language and discourse (option sets and solutions that were previously marginal gain

prominence, reflecting a re-weighting of choices society deems viable), and coordinated behavioral tilts — such as masses of people engaging in preventative health measures, seeking reconciliations (making amends, forgiving debts/offenses), or generally de-escalating conflicts during times of shared crisis. These would be population-level analogs of individual horizon effects: as the “shadow of the future” grows short for many, we collectively prioritize actions that keep the overall ledger balanced and de-prioritize those that could cause irreparable harm.

In summary, the Law of Fairness doesn’t just apply to single life streams in isolation; it may also manifest in how societies respond when a large number of lives approach critical horizons. A rising population shadow price Λ_t would mean society as a whole is feeling the squeeze of time and is moved to enact “last-minute” balancing operations (through policies and cultural shifts) to preserve fairness across the board. These policy windows are where we should look for real-world evidence of the theory, watching if our institutions and communities indeed behave in ways that mirror the urgency of a closing horizon for many people at once.

6.7.1 Where we go next:

With individual and population levers in view, Part IV turns to measurement itself. Chapter 7 builds the Hedonic Composite Index (HCI) so we can track experience in the right units; Chapter 8 asks how to keep those units aligned across people, places, and time.

Part IV — Measuring Feeling Without Fooling Ourselves

Have you ever wondered how we prove that a system is fair? It's one thing to declare lofty principles, but it's quite another to measure fairness in everyday life. How would we actually know if each person is getting a fair share of comfort and relief? In Part IV, we turn the Law of Fairness (LoF) into concrete numbers that we can track. We'll talk about metrics – the gauges and yardsticks that tell us if LoF's guarantees are being met. In this Part, "guarantee" means a testable constraint: if preregistered measurement repeatedly fails to detect it, we treat that as evidence against LoF in that domain. After all, under LoF the system must do more than just *intend* fairness; it must demonstrably deliver fairness through measurable outcomes. This means breaking down abstract ideas like comfort, relief, and dignity into data points we can observe over time.

Think of it like tracking your fitness: you might count steps, monitor heart rate, or log meals. Here, we're tracking "*fairness fitness*." We need a way to quantify an individual's well-being (so we know when someone needs more help) and a way to keep score of what each person has received over their journey (so we can correct any imbalances as they arise). To do that, we'll build a Hedonic Composite Index (HCI) to capture a person's day-to-day comfort level. We'll also use personal ledgers to sum up those comfort "units" over time, much like a bank account accrues funds. Periodically, the system will check neutrality gates – special checkpoints (for example, at end-of-life) where everyone's "account balance" of comfort should line up within strict tolerances. Because $\hat{L}(t)$ is estimated, neutrality is assessed with uncertainty: gates pass when ledger differences fall within the tolerance band under the preregistered equivalence gates given the credible interval. All of these measurements feed into the system's decision-making. In fact, the LoF algorithm's action function at any given time—written as $A(t; \hat{L}, H_t, C)$ —takes these values as inputs: HCI_t (or latent F_t) for a person's current hedonic state, H_t for time-horizon, C for context, and $\hat{L}(t)$ for the running estimated ledger of relief received (with uncertainty bands). Because later chapters use H_t for time-horizon, we'll write the hedonic-state input as the current HCI_t (or latent F_t) and reserve H_t for horizon when both appear. $\hat{L}(t)$ denotes the system's running estimate of the ledger (with uncertainty), while $L(t)$ denotes the underlying latent ledger. With the right metrics in place, $A(t; \hat{L}, H_t, C)$ can determine who needs what and when to keep things fair.

Equally important, we'll ensure our metrics are both meaningful and ethical. A number is only useful if it represents the same thing for everyone – otherwise, comparisons become apples-to-oranges. We'll test something called measurement invariance, checking that our indices (like HCI) truly measure the *same construct* for different people (so an HCI of 50 means the same level of comfort whomever it's measured on). If

invariance fails, we do not treat equal numbers as equal comfort across groups; we restrict to within-person change or report group-specific scales instead. And while we plan to measure everything that matters, we won't forget the human element: "*Relief is a systems variable; comfort and dignity override data collection.*" In practice, this means if a patient is in pain or distress, the system prioritizes easing that suffering over poking or prodding for more data. Our goal is fairness with compassion – metrics that inform and improve care, not metrics for metrics' sake.

What this Part will do for you:

- Provide a clear understanding of why measuring fairness is essential, and how LoF turns fairness into guarantees we can test rather than just hopeful intentions.
- Introduce the Hedonic Composite Index (HCI), a plain-language gauge of a person's comfort and relief at any moment (the core of our fairness metrics).
- Tracking fairness over time: How the system uses units of relief and personal ledgers to track well-being over time – ensuring that each individual's accumulated comfort (represented by $\bar{L}(t) = \int_0^t HCl(\tau) d\tau$) stays on course.
- Provide an explanation of neutrality gates (like an end-of-life checkpoint) where LoF requires outcomes to align within very tight margins – ensuring that by critical moments, no one is left far behind in comfort or dignity.
- Tools and validation: An overview of the statistical tools we use to validate these metrics (from simple counts modeled by Poisson to more advanced model checks like the Widely Applicable Information Criterion, WAIC, for predictive accuracy), all while keeping ethical guardrails firmly in place.

Chapters in this Part:

- **Chapter 7 — The Hedonic Composite Index (HCI):** Builds a joint metric of felt experience from multiple channels, lays out blinding/prereg rules, introduces latent-variable/state-space tools, defines Hedonic Composite Units (HCU), and lists fail conditions.
- **Chapter 8 — “Same Scale” Across People and Places:** Tests whether HCI means the same thing across groups and contexts, tightens invariance, and shows how to compare and pool responsibly without erasing difference.

Where we go next:

Turn the page to Chapter 7. We'll assemble the composite, stress-test it with blinds and preregistration, and define the unit we'll carry into the ledger. Measurement comes first; claims follow.

Chapter 7 — The Hedonic Composite Index (HCl)

Picture a busy hospital ward late at night. A nurse making rounds has to decide which patient needs attention first. One patient grimaces quietly, rating their pain a “5 out of 10”; down the hall, another moans loudly yet also calls their pain a “5.” The nurse knows these 5s might not mean the same thing – one patient could be stoically underreporting agony while the other exaggerates discomfort. How do we put a number on something as personal as comfort or relief? In this chapter, we set out to do exactly that. The Hedonic Composite Index (HCl) is our answer to a big challenge: creating a single score that reflects a person’s overall well-being at a given moment. If you’re responsible for ensuring each patient in a care facility feels as comfortable as possible, you can’t just ask “Are we fair?” – you need data. HCl is like a vital sign for fairness: a simple, easy-to-understand number that summarizes how someone is doing in hedonic terms—i.e., in terms of pleasure, pain, comfort, and distress. Formally, HCl_t is our uncertainty-aware estimate of the momentary net-affect variable F_t that will later be integrated into the ledger.

First, we’ll unpack the logic behind HCl. The idea is that no single factor captures well-being. Pain levels, emotional state, energy, anxiety, even dignity or morale – all contribute to how “good” or “bad” someone feels. HCl combines multiple inputs into one composite metric. In plain language, think of it like a recipe: we take a bit of this (perhaps a pain score), a bit of that (a mood rating, a mobility level, etc.), and blend them with appropriate weights that reflect their importance. The result is one score that rises when things are going better and falls when things are worse. This single score gives the LoF system a clear target to monitor and maintain. Under LoF, falling HCl should be accompanied by stronger compensatory pressures (e.g., relief-seeking or acceptance). Operationally, that “pressure” is a shift in predicted choice/behavior (e.g., greater weight on high-Φ reparative options), not evidence of purpose. Our task is to measure these shifts without implying purpose or moral desert.

How do we decide on the ingredients and weights of HCl? This is where data and method come in. We’ll describe how we selected a set of indicators that together cover the full hedonic experience – for example, by using validated pain scales, comfort surveys, and observational metrics. Each component is standardized (so they’re on the same numerical footing) and then combined. We didn’t just eyeball this; we relied on statistical modeling to ensure HCl is robust and reliable. Different candidate models were tested (imagine trying out slightly different “recipes” for combining inputs), and we chose the version that best predicted real outcomes while aligning with expert judgment. To keep ourselves honest, we used modern out-of-sample validation — for instance, computing the Widely Applicable Information Criterion (WAIC) for each candidate model. WAIC (a

Bayesian cousin of the classic AIC) estimates how well the HCI model would predict new data while penalizing over-complexity. We treat small WAIC differences as inconclusive unless corroborated by held-out predictive performance. In short, the model with the lowest WAIC gave us a composite that wasn't overfit and generalized well. (We name WAIC here once; details of model comparison appear in the Research Notes.)

By the end of the chapter, HCI will go from an abstract acronym to a useful mental tool. We'll walk through example scenarios: for instance, a patient's HCI might be 75 (on a 0–100 scale) on a good day and 40 on a bad day. When we switch to 0–1 examples (e.g., 0.8), it's just this same scale normalized (e.g., 80 on a 0–100 scale). The higher the HCI, the more hedonic "goodness" that person is experiencing. This single number lets the system compare needs at a glance and allocate help accordingly. It also feeds into the personal ledger (covered in the next chapter) via $\bar{L}(t) = \int_0^t \text{HCI}(\tau) d\tau$, meaning we accumulate HCI over time to track the total comfort a person has experienced. One neat analogy: this is similar to the idea of quality-adjusted life years (QALYs) in healthcare, where you integrate quality of life over time. This is an analogy for the integration arithmetic (area under a quality curve), not a claim that HCI is identical to QALY utility weights. If someone averages 0.8 on a normalized HCI scale for a full year, they accrue roughly 0.8 "hedonic-years" of comfort in that year. Another person averaging 0.4 would accrue 0.4 – only half as much comfort – indicating a fairness gap that LoF would not allow to persist indefinitely. HCI gives us the moment-to-moment readings needed to detect and then close those gaps before they grow too large.

What you'll get from this Chapter:

- Why a composite index? Understand the rationale for HCI – why fairness can't be measured by any single factor like "pain score" alone, and how a composite paints a richer, more reliable picture of well-being.
- What goes into HCI: A plain-language breakdown of HCI's ingredients – from physical comfort indicators to emotional and cognitive measures – and how they're combined into one meaningful score.
- How HCI is built and validated: Insight into the methodology for weighting components (e.g. using statistical techniques like factor analysis or regression) and how we chose the best model by checking which composite best predicted outcomes (using a rigorous criterion like WAIC to avoid overfitting).
- HCI's role in LoF's system: See how HCI is used in real time by the LoF algorithm. The system continuously monitors each person's HCI and uses it to decide when to intervene. A persistently low HCI triggers an action $A(t; \bar{L}, H_t, C)$ to boost that individual's comfort, ensuring no one's index stays low for long.

- Trust and consistency: Learn how we verified the HCI's reliability and consistency – for example, checking internal consistency and making sure that two people with the same HCI truly have comparable levels of comfort. (This is a prelude to the invariance tests we'll tackle in Chapter 8.)

Subsections in this Chapter:

- **7.1 Inputs: Report, Physiology, Brain, Behavior, Dreams** - The five inputs and why each adds unique information, with minimal, humane collection plans.
- **7.2 Why Composite Beats Single Meters** - How joint models reduce error and bias, and when a single channel misleads while the bundle stays honest.
- **7.3 Keeping It Honest: Blinds and Preregistration** - Concrete blinding targets, prereg templates, and adversarial checks that keep us from convincing ourselves.
- **7.4 Research Notes: Latent (CFA/IRT) and State-Space** - The modeling layer that turns noisy channels into a coherent latent trajectory, with stop rules for nulls.
- **7.5 Hedonic Composite Units (HCU)** - Defining the unit so ledgers can add up meaningfully, with behavioral and psychophysical anchors.
- **7.6 Fail Conditions for HCI** - The ways measurement can and should fail; if these occur, we downgrade or rebuild the index rather than defend it.

Where we go next:

Section 7.1 inventories the inputs. Keep that list close; every modeling choice in later sections must respect what each channel can and cannot tell us.

7.1 Inputs: Report, Physiology, Brain, Behavior, Dreams

HCI treats affect as a latent state revealed by five imperfect witnesses. Each witness brings unique variance and unique failure modes. For each input channel, we specify *what* is collected, *how* it is cleaned, *how* it enters the model, and what would disqualify it from contributing.

7.1.1 Self-report (micro-EMA and structured prompts)

What we collect:

- Micro-EMA surveys – Frequent, very brief mood surveys (1–3 items on 5–7 point scales) delivered at random times and at key events (e.g. just before and after significant actions).
- Sliders for affect – A continuous rating slider from “very unpleasant” to “very pleasant,” a load/relief slider for how burdensome or relieving the moment feels, and a somatic marker tag (choosing a descriptor like “heavy/tight” vs. “light/open” to characterize bodily feeling).
- One-line context – A free-text field for notes or context, which will later be converted into privacy-preserving embeddings (no raw text leaves the device).

Preprocessing:

- Responses are z-scored per person (each person serves as their own baseline), and corrected for diurnal trends (time-of-day effects). This within-person z-scoring supports tracking changes over time; cross-person comparisons rely on the multi-group measurement model and HCU anchoring rather than raw z-scores. Sparse missing responses are imputed using last-observation-carried-forward and state-space smoothing techniques. Last-observation-carried-forward is used only as a temporary initialization; primary inference relies on the state-space model’s missing-data handling with propagated uncertainty.
- Language normalization: Multilingual prompt sets are used for cross-cultural deployments, with back-translation checks to ensure semantic consistency of items across languages.

Model entry:

- Self-report enters HCI as a linear “loading” on the latent state F_t . For example, a simplified model might be: $y_t^{\{EMA\}} = a_{\{EMA\}} F_t + b^T Z_t + \varepsilon$, where Z_t includes nuisance covariates like arousal and effort. (In practice, we use an ordinal IRT model for Likert items – see 7.4.1.)

- Anchoring: Controlled shifts such as analgesia (pain relief) and cold-pressor pain (see HCU anchors in 7.5) are used to calibrate this self-report slope and intercept, so that a unit change has consistent meaning across studies.
- Field instrumentation. Modern wearables and phones provide passive, privacy-preserving proxies for $\text{HCI}\Delta$: heart-rate variability and sleep staging from wearables; mobility dispersion and routine stability from GPS; speech prosody features (on-device); app/communication rhythms. Here $\text{HCI}\Delta$ refers to per-interval change signals (ΔHCI) inferred from passive features, not a separate index. We preregister a sensor feature set, compute person-standardized $\Delta z_i(t)$, and integrate to $\hat{L}(t)$ with block-bootstrap uncertainty. This enables long-horizon tests without heavy respondent burden.

Quality / exclusion:

- If a participant “straightlines” (gives the same answer every time) or responds impossibly fast, those entries are down-weighted. Any individual with <50% compliance (failing to answer at least half of prompts) gets their self-report channel excluded from HCI for that study.
- Invariance checks: We probe configural→metric→scalar invariance for the self-report items across languages and cultures (see Chapter 8). If the meaning of ratings doesn’t hold (e.g. different cultures use the scales differently), we either adjust or restrict cross-group comparisons.

Why it matters: Self-reports directly access qualia – the person’s own description of their experience – providing face validity. They also set the anchors for other channels, giving a common reference for what “+1 unit of relief” feels like in subjective terms.

7.1.2 Physiology (autonomics and sleep)

What we collect:

- Cardiac signals: Heart rate (HR) and heart rate variability metrics (e.g. RMSSD, high-frequency HRV) from ECG or PPG sensors.
- Electrodermal and ocular: Skin conductance level and response (SCL/SCR) for sweating, pupil diameter for arousal, respiration rate/variability, and peripheral skin temperature.
- Sleep metrics: When feasible, full polysomnography (PSG). Otherwise, validated wearables for sleep, tracking things like REM density (as a proxy for intense dreaming) and slow-wave sleep power (deep restorative sleep).

Preprocessing:

- Artifact rejection: Remove movement artifacts (e.g. toss out HRV segments with excessive movement or ectopic heartbeats), interpolate blink artifacts in pupil signals, etc.
- Context corrections: Add temperature and season as covariates (to account for sweat and heart rate baseline shifts in hot vs. cold climates), and apply per-device calibration curves if different devices are used (ensuring two different wristbands give comparable readings).

Model entry:

- Multivariate observation: Autonomic readings often enter as a vector. For example, we might model $\mathbf{y}_t^{\text{auto}} = \mathbf{A}_{\text{auto}} \mathbf{F}_t + \mathbf{B}_{\text{auto}}^T \mathbf{Z}_t + \varepsilon$, where $\mathbf{y}^{\text{auto}}_t$ could include HRV, SCL, pupil, etc., and \mathbf{A}_{auto} is a set of loadings.
- Arousal control: Importantly, arousal-related factors (like pupil size, which is both an affect signal and an arousal signal) live in the nuisance vector \mathbf{Z}_t . This ensures that autonomic contributions to HCl aren't just tracking raw arousal.
- Sleep dynamics: We include specialized terms: entering slow-wave sleep (SWS) contributes a variance-reduction effect (people's affect volatility typically drops in deep sleep), and REM sleep contributes a "counterweight capacity" term that will gate dream inputs (see 7.1.5 on dreams). These SWS/REM terms are treated as testable modeling hypotheses and are retained only if they improve preregistered out-of-sample prediction; we don't assume the effect a priori.

Quality / exclusion:

- Specific channel drop rules: If an HRV series is unusable (e.g. >25% of beats are ectopic), we drop that HRV measure for that person. Skin conductance data taken while a topical agent (e.g. antiperspirant or medical cream) is on the skin is excluded. Pupil tracking is excluded if the participant is on mydriatic drugs that dilate pupils.
- If a wearable device is swapped out for a different model without a fresh calibration, the sleep data from that transition is excluded (device-specific bias could otherwise masquerade as a change in the person).

Why it matters: Physiological signals provide an embodied readout of how easy or heavy experience is, often preceding conscious report or deliberate action. For instance, heart rate variability might dip before someone even articulates feeling stressed. These signals can reveal hidden fluctuations in affect and arousal that self-report or behavior might miss.

7.1.3 Brain (fMRI/EEG/MEG signatures)

What we collect:

- fMRI regions: Key ROI (Regions of Interest) in the brain associated with value and control. For example: ventromedial prefrontal cortex (vmPFC)/orbitofrontal cortex (OFC) for reward value, anterior cingulate cortex (ACC) for cost or pain signals, right inferior frontal gyrus (rIFG) and subthalamic nucleus (STN vicinity) for inhibitory control (“brakes”), anterior insula for interoceptive awareness. (Optionally, we may monitor hippocampus, dorsomedial PFC, or TPJ for memory and social cognition aspects.) These ROI/function mappings are coarse and task-dependent; we treat them as preregistered hypotheses and interpret them as correlates, not direct readouts of subjective feeling.
- EEG/MEG markers: We extract well-known electrophysiological features: frontal midline theta (as an ACC proxy for conflict/effort), motor beta suppression (reflecting rIFG/STN braking of actions), the late positive potential (LPP) for emotional significance/meaning, and the contingent negative variation (CNV) or CPP which reflects urgency as a deadline approaches.
- Tasks to evoke signals: Participants may perform utility-matched choices that differ only in compensability features (e.g. choosing a “Repair” option vs. an indulgence that have equal immediate reward), with horizon instruments (like explicit countdown timers) and congestion manipulations (like visibly differing queue lengths) embedded in the task. These ensure brain signals are probed under scenarios where LoF dynamics should appear.

Preprocessing:

- For fMRI: apply standard motion correction and physiological noise regressors (e.g. RETROICOR to account for heartbeat and respiration effects if available). Spatial smoothing and high-pass filtering per preregistered pipeline.
- For EEG/MEG: perform ICA to remove ocular (blink) and muscle artifacts; apply only the pre-specified filters and time–frequency windows that were preregistered (so we don’t go fishing for a nice-looking band post hoc).

Model entry:

- Brain as dual evidence: We treat processed brain signals as both direct observations of F_t and as auxiliary checks on LoF mechanisms. For fMRI, we extract trial-wise beta weights from ROI GLMs (general linear models) — for instance, how strongly did vmPFC activate for a given choice — and use those betas as inputs indicating affect F_t . We also check whether those betas scale with

- LoF quantities (e.g. does rIFG activity covary with a $\Phi \times H^{-1}$ term under preregistered coding and contrasts?). Braking predictions are about imbalance risk: with shorter horizons, inhibitory-control signatures should strengthen for low-compensability options; depending on how Φ is coded, that may correspond to either sign in a $\Phi \times H^{-1}$ term.
- For EEG, we condense each trial's data into summary amplitudes (theta power, beta rebound, LPP amplitude, etc.) and let those load on F_t (again with nuisance controls, such as overall arousal level, included to isolate the affect-specific component).

Quality / exclusion:

- Motion censorship (fMRI): If a run has excessive motion (e.g. mean framewise displacement > 2 mm), we either exclude that run or include “censor” regressors for the bad time points so they don't contaminate the signals of interest.
- EEG trial loss: If an EEG session retains less than ~70% of trials after artifact rejection, we throw out that session's EEG data (too much noise to trust). Also, any EEG channel that shows abnormally high impedance throughout (poor contact) is dropped from analysis upfront.

Why it matters: Brain measures give us mechanistic leverage. They constrain mechanisms only to the extent that the preregistered tasks and contrasts isolate them; on their own, these signals remain indirect and correlational. They can show that the latent affect F_t is not just an abstract number, but corresponds to known neural dynamics: for example, vmPFC “value signals” and rIFG/ACC “braking signals” should covary with F_t and specifically with the horizon-related scaling (short horizons \rightarrow bigger differences) if LoF is correct. These neural signatures help confirm that HCl isn't just picking up generic emotion or stress, but the specific pattern predicted by the theory.

7.1.4 Behavior (micro-choices, stickiness, policy stall)

What we collect:

- Decision outcomes: Choices among options that have been matched for immediate utility but differ in LoF-relevant ways (e.g. choosing a small *reparable* harm vs. a small *irreversible* harm). Reaction times (RT) for decisions. Whether the person aborts or persists in multi-step goal sequences when given the chance to quit early.
- Naturalistic “stickiness”: Outside the lab, we track things like: do people finish what they start? (e.g. drafts written vs. discarded emails, phone calls placed vs. abandoned mid-dial). Social contact patterns can be features too: proportion of

- outreach that is reparative (seeking comfort) vs. adversarial on high-strain days. We also consider repertoire size per day (how many distinct activities or social contacts a person engages in – this often shrinks under heavy load).
- Motor inhibition tasks: Classic lab tasks like the Stop-Signal Task or Go/No-Go, which provide metrics (stop-signal reaction time, commission errors) indicating how much control (inhibition) the person is exerting. These tie into affect via frustration or effort in the face of constraints.

Preprocessing:

- Outlier removal: Extremely fast reactions (<200 ms) or implausibly slow ones (>3 SD from a person's mean) are removed to avoid skewing the analysis.
- Latent parameter estimation: We may fit models like the drift-diffusion model to decision data to extract latent parameters (e.g. caution levels, implicit bias toward quitting). This is done hierarchically so that we get stable estimates per person.
- Privacy for text data: If we use text (e.g. the content of emails or messages to classify them as positive/negative), we do so via on-device embedding models. Only summary features (like “contact valence score”) leave the device, and the raw text is not stored centrally.

Model entry:

- Direct effect on F_t : Certain behaviors feed back into the latent affect state. In the model this is implemented as a timed input/event term; causal direction is tested with interventions and timing, not assumed from observational association. For example, if someone performs a clearly relieving action (like finally fixing a problem), we often see their physiology settle; this kind of event can be modeled as nudging F_t upward (less burden).
- Covariates for tests: Behavior also provides outcome variables for testing LoF predictions. For instance, when modeling the probability of choosing an option u , we include not just utility but the option's $\Phi(u)$ (compensatory potential), the interaction $\Phi(u) \times H^{-1}$ (is this extra appealing when time is short?), and any social penalty terms $\sum_r \lambda_{rt} \Delta_r(u)$ if the choice uses up shared resources. By including HCl (current F_t) as a predictor too, we test if being in a certain affective state biases choice as predicted.
- Derived measures: “Stickiness” (whether you stick with a plan) and abort rates are treated as secondary observables that should shift when the admissible set changes. For example, if horizons shrink and unrewarding plans should be

pruned, we expect more aborts on low-value tasks; HCI dynamics might predict those changes.

Quality / exclusion:

- If a participant did not find an experimental manipulation credible (say, they didn't believe the time pressure was real), we exclude that task's data from contributing to HCI training, though we might still analyze it separately for insight. We require that horizon and congestion instruments pass their manipulation checks (the participant noticed them and reacted appropriately in manipulation-specific questions) or else that dataset is flagged in analysis.

Why it matters: Behavior is where the “rubber meets the road” – it converts silent internal tilts into externally auditable actions. Did the person actually slow down or speed up? Quit or persevere? Shift toward comfort-seeking? These are observable consequences of the internal state. If HCI is meaningful, it will correlate with and predict these real-world actions in the patterns LoF specifies (e.g. disproportionate quitting of low-compensability tasks when time is short, etc.). Behavioral signatures keep our index honest: if HCI says someone is deeply burdened but none of their choices or performance measures change, we've likely got a problem with the index.

7.1.5 Dreams (night-shift counterweights)

What we collect:

- Morning dream reports: Upon waking, participants give a simple affect rating for their dominant dream (e.g. how positive or negative) and tag any major themes with short labels (did the dream involve repair/healing, confrontation, loss, mastery, etc.).
- Lucidity and recall: We note if the dream was lucid or not, and we record a minimal narrative length or detail to gauge recall quality. We also log the prior day's net strain from HCI (ΔL_{day}) to relate to the dream.
- EEG-triggered sampling (if possible): In some cases, if the participant has PSG, we attempt to capture affect during REM via brief micro-awakenings (waking them up momentarily to report feeling, then letting them continue sleep). This yields a more immediate dream-affect integral but is only done with full consent and in research sleep labs.

Preprocessing:

- Normalize each person to their typical recall and affect reporting style (some people never remember dreams; others always do). Control for any medications

(e.g. SSRIs, benzodiazepines) and factors like alcohol or sleep debt that can suppress REM or alter dream affect.

- Free-text descriptions (if any) are encoded into affective feature vectors using the same privacy-preserving approach (local embedding) as other text; again, no raw narrative leaves the device.

Model entry:

- Dream affect as an input: We compute a dream-affect integral D_n for each night (based on the intensity and valence of dream emotions). This enters the model as a nocturnal input that is hypothesized to *invert* the previous day's drift in the ledger. The explicit null is D_n independent of ΔL_{day} and H (i.e., no systematic inversion and no horizon scaling). In formula: $D_n \approx -\alpha(H) \Delta L_{\text{day}}$, with $\alpha(H)$ (the “counterweight gain”) expected to increase as horizons shorten. In plain terms: the closer a person is to an ending, the more strongly a night's dreams should push back against the prior day's emotional imbalance.
- Additionally, D_n is allowed to modulate the next day's admissible set ease. For example, a very reparative dream might effectively refresh one's willingness to engage, which we model as lowering the subjective cost of reparative actions come morning.

Quality / exclusion:

- If a night has unusual sleep architecture (e.g. REM latency > 3 hours, or REM completely suppressed due to some disturbance or drug), we flag those data points – the absence of REM or extreme delay might indicate the brain didn't do its usual processing. On no-dream-report nights, we treat D_n as zero with a large uncertainty (since we don't know if nothing happened or just nothing was recalled).
- Opt-out protection: Because dream content can be highly personal, participants can opt out of sharing dream data at any time without penalty. HCI can run without the dream channel; it will simply carry larger uncertainty bands to reflect the missing counterweight input.

Why it matters: Dreams provide a low-cost compensatory mechanism – a way for the brain to rebalance affect without external action. We treat this as a falsifiable claim about measurable covariation with next-day affect and behavior, not as a claim that dream narratives have fixed meanings. The theory predicts that as horizons shorten, this mechanism (dream “relief”) should strengthen. By including dreams, HCI can capture

some of that overnight emotional processing. It's an unusual input for a psychometric index, but if LoF is real, dreams are not epiphenomenal; they're part of the balancing act.

7.1.6 Nuisance and context covariates (never optional)

To avoid circularity and false positives, every observation channel is paired with nuisance covariates in the model. We always record metrics for:

- Arousal: e.g. pupil size, electrodermal activity (these often indicate general arousal or stress level).
- Effort: e.g. physical activity or reaction time changes that reflect effort.
- Conflict: e.g. entropy of choices or conflict in decision tasks.
- Fatigue/sleep debt: recent sleep history, time awake.
- Medications: especially psychoactive meds (antidepressants, anxiolytics, stimulants).
- Circadian/seasonal context: time of day, ambient light; season or temperature for physiological baselines.
- Movement: for imaging sessions, motion parameters.
- Trait baselines: individual differences like baseline depression or anxiety scores, if relevant.

These Z_t covariates are never optional – if a purported LoF effect disappears when a justified nuisance variable is added, we count that as information, not failure. (We want a rival explanation to win if it should win; we are not in the business of hiding confounds. For example, if a “horizon effect” vanishes once you control for stress hormones, then perhaps it was just stress all along, and we need to redesign the instrument.)

7.1.7 Fusion and weighting (how inputs combine)

HCI is built via state-space fusion, meaning it dynamically weighs each channel by its reliability. Each channel K gets an estimated loading a_k and noise level σ_k ; these are often modeled hierarchically (varying by person around a group mean). The fusion algorithm (typically a Kalman filter/smooth or particle filter) will trust channels less when they're noisy or inconsistent. For example, if someone's wearable HRV device starts acting up (higher noise variance), that channel's influence on F_t automatically down-weights.

We also implement channel holdouts as a diagnostic: in K -fold fashion, we periodically rebuild HCI leaving one channel out, and then see if that channel's observations can be predicted from the others out-of-sample. If, say, the EEG channel is completely unpredictable from the rest and also doesn't add predictive power elsewhere, it may be contributing nothing but noise – a candidate to drop.

Crucially, we define an “ethical minimum” sensor set: by default, HCI is computed with self-report + HRV + pupil only. These are relatively low-burden and non-invasive. Brain and dream inputs are optional add-ons that must earn their keep via incremental predictive value *and* require explicit consent. This ensures we never say “you must wear an EEG cap to measure well-being” unless it’s truly justified.

7.1.8 Exclusion, down-weighting, and fail flags

We build in rules to catch data quality issues and fail-fast if something threatens validity:

- Device swap: If a person switches to a new device (e.g. a different brand of fitness tracker) without a calibration period, we immediately down-weight or omit that physiology data until we recalibrate.
- Language mapping issues: If our language model for text sentiment doesn’t perform well in a certain language/dialect (high uncertainty or weird mappings), we exclude those free-text embeddings and stick to numeric EMA responses.
- Excessive motion in brain data: Rather than aggressively “denoising” data with heavy artifacts (which risks inventing signal), we prefer to censor or exclude sessions with high motion, as noted above.
- Medication confounds: If someone starts a medication mid-study that we know affects mood or physiology, we either include an interaction term or exclude that segment of data. We won’t ignorantly average through a medication effect.

We also set channel-level fail tests:

- If adding a channel does not improve held-out prediction of key outcomes (choices, autonomic responses, or other channels’ values), then for that cohort we drop the channel from the composite. In other words, each channel must earn its place by contributing something unique.
- If two or more major channels (say, both self-report and HRV) fail invariance or predictive checks at a given site, we will not compute a single HCI at that site at all—we’d report a site-level null result. This prevents us from cherry-picking partial results; if the composite can’t be trusted in one location, we don’t quietly run it with half the pieces and pretend everything is fine.

7.1.9 Privacy, consent, and data minimization

Ethical data practice is built into HCI:

- Local feature extraction: All raw text, audio, or high-dimensional personal data (like full diaries or voice recordings) are processed locally on the participant’s device. Only abstracted features (scores, counts, embeddings) are sent to

researchers. For example, we might store “daily sentiment = 0.2 (mild positive)” rather than the actual journal entry.

- Granular consent: Participants choose which channels they’re comfortable with. Someone might agree to wear a heart-rate monitor and answer EMAs, but not to share dream reports; that’s fine. HCI will degrade gracefully, simply widening the uncertainty bands when channels are omitted. Each channel’s contribution is designed to be modular.
- Right to withdraw: Participants can withdraw from the study (or ask for their data deletion) at any time. If they do, their data and any derivatives are purged or anonymized, and this deletion propagates to any aggregated results where feasible. We build pipelines that can re-run analyses excluding a given participant if needed.

7.1.10 What success looks like (per channel)

Finally, we have pre-registered channel-specific success criteria – patterns each channel *should* exhibit if HCI and LoF are on the right track:

- Self-report → others: Self-reports should be predictive of upcoming physiology and behavior via HCI. Importantly, it shouldn’t be only one-way. If all we see is that physiology predicts self-report but not vice-versa, then self-report might be redundant or lagging. We expect that if your HCI (which heavily weights self-report here) is high now, your body and choices in the next hour or day will reflect that (calmer physiology, constructive actions). Not just the other way around.
- Physiology settling: Physiological indicators (like HRV recovery) should settle faster after high-Φ choices when horizons are short. In plain terms: as the end looms, if you do something reparative or relieving, your body should calm down more quickly than it otherwise would – a sign that the choice “hit the spot” in balancing the ledger.
- Brain signatures: In neural data, we expect to see vmPFC valuation increases and rIFG/ACC braking signals that covary with the latent affect F_t and, specifically, with horizon interactions ($\Phi \times 1/H$). For example, as H shrinks, rIFG should kick in harder (more braking) for any action that would cause imbalance, and vmPFC should show extra “reward” activation for truly reparative options. If the brain isn’t showing these patterns, then HCI’s improvements might be capturing something else.
- Behavior biases: Behavior should show stickiness asymmetry – e.g. after a high-HCI day (lots of relief, positive experience) versus a low-HCI day (painful experience), the next day’s engagement patterns might differ. Also, policy stall:

when HCI dynamics indicate a person is in a burdened state, we expect to see them stall on multi-step risky plans unless those plans are clearly compensable. In short, the choices people make and whether they persist should tie back to HCI, above and beyond utility considerations alone.

- Dream inversion: Dreams, as noted, should invert prior-day drift – the worse the day's imbalance, the stronger the dream rebound – and this effect should be *proportional to horizon*. If someone's near the end of a critical period (short horizon) and had a strongly negative day, their dream rebound should be especially pronounced.

We've pre-registered these as checks. If two or more channels repeatedly fail to show their expected pattern (despite adequate power and clean instruments), our policy is to publish the null result and revisit the composite. That is, we don't just quietly ignore it – we either fix HCI or refine the theory.

7.1.11 Where we go next:

With inputs on the table, 7.2 explains why a bundle beats any single meter—how joint evidence reduces error, catches drift, and preserves meaning when one channel goes noisy.

7.2 Why Composite Beat Single Meters

A single dial cannot capture the weather. Likewise, no single measure – whether a mood scale, HRV reading, pupil size, vmPFC activation, or dream affect – can stand in for felt experience across all contexts, cultures, and times. A composite wins for principled reasons that matter scientifically and ethically.

7.2.1 Different channels carry different truths (and different lies)

- Self-report taps directly into subjective feeling, but can bend under biases: demand characteristics (wanting to please the researcher), language nuances, stigma or cultural norms about admitting certain feelings.
- Autonomic signals reflect embodied ease vs. strain, but they confound with arousal and other factors: a hot room or exercise can spike your heart rate or sweat responses, unrelated to mood; posture changes or caffeine can affect these too.
- Brain ROI signals reveal mechanistic details (like “the reward center lit up”), but brain measures are costly, sensitive, and narrow: fMRI demands you lie in a noisy tube, EEG is prone to motion artifacts, and any brain result may depend on the specific task you were doing. We avoid reverse inference: ROI phrases are shorthand, and the evidence lives in preregistered patterns and cross-channel convergence—not in a single region “meaning” one thing.
- Behavior is externally verifiable (we can all agree if someone pressed a button or quit a game), but it’s highly context-bound: a hesitant pause means different things in a job interview vs. at home, and policy choices depend on what’s available – behavior alone may miss internal states if people are coping.
- Dreams expose subconscious processing – they can reveal conflict resolution or lingering distress overnight – but recall varies wildly and medications or substances can suppress dreaming entirely.

A composite fuses these complementary signals while letting the independent noise cancel out. No single channel gets to “lie” without the others calling it out. If self-report says “all good” but physiology and behavior scream “stress,” HCl will reflect the discrepancy as uncertainty or conflict. If all channels agree, we gain confidence that we’re measuring the true latent state.

7.2.2 Missingness and asynchrony are the norm

In real life, data will be incomplete: you skip survey prompts, your wearable battery dies, you only get in the scanner once a month, and you forget most of your dreams. State-space fusion in a composite gracefully handles this. It can take whatever channels are

present at a given moment to update F_t , and carry forward uncertainty during gaps. The result is a continuous estimate \hat{F}_t with credible intervals that widen when data are sparse. A single-meter approach has to either drop those days (losing information and biasing samples) or fill them in with guesses (“Last value carried forward” – which can become wishful imputations). With HCl, missing data just means more reliance on the state model and other channels; the uncertainty tells us how much to trust it.

7.2.3 Invariance is easier to earn jointly than alone

No single meter is culturally neutral. Self-reports shift with language and local norms (“7/10” means different things in different places); autonomic baselines shift with climate or ethnicity (skin conductance in humid vs. arid climates, for example); behavior norms differ by policy (pain expression in stoic vs. expressive cultures). But a composite can recalibrate loadings per site while constraining the overall structure. For instance, maybe in one culture the self-report has a lower loading and physiology a higher loading – HCl’s multi-group model can adjust that (metric invariance) yet still maintain a common latent scale. By achieving invariance at the latent level (after composite fusion), we can compare apples to apples even if any one channel was apples-to-oranges across groups. In short, joint calibration beats trying to force a single measure to work everywhere.

7.2.4 Causal identification needs converging consequences

If we perturb true affect, multiple outputs should echo the change. Say we give fast-acting analgesia (pain relief). Then according to LoF: HCl should go up (more relief), and we’d expect to see report of relief ↑, HF-HRV (often interpreted as a relaxation indicator) ↑, vmPFC activity (often interpreted as a value signal) ↑, rIFG “braking” ↓ (often associated with inhibitory control; because there’s less need to suppress pain-driven responses), and fewer aborts of tasks (since pain is relieved). A single channel can’t capture this multi-faceted consequence and, crucially, can’t discriminate cause from confounds: if only self-report moved, was it really pain relief or placebo effect? If only heart rate changed, was it just sedation? A composite, by demanding a pattern across report+body+brain+behavior, is far more specific. It enforces pattern consistency that helps strengthen causal interpretation of interventions.

7.2.5 Out-of-sample prediction is the real referee

Practically, we judge models by how well they predict new data. Over many participants and days, composite HCl consistently improves held-out predictions of key outcomes, whereas single metrics often generalize poorly. We only count a “predictive win” when it appears under preregistered, blocked validation (and we report the score differences

with uncertainty), not from in-sample fit. Concretely, our studies show that using HCI (vs. any single channel) gives better forecasts of:

- Choice behavior among options matched for immediate reward (the composite helps predict which option people will pick when those options differ in long-term compensability Φ or in whether they provide relief). It's especially useful through terms like Φ and $\Phi \times H^{-1}$ in choice models.
- Stickiness vs. stall in multi-step plans (who gives up vs. who persists, under what conditions).
- Autonomic settling after commitments (how quickly heart rate or skin conductance returns to baseline after a major decision).
- Brain ROI modulation in key regions (how strongly vmPFC, rIFG, ACC, etc., respond during tasks).

Single-channel measures rarely predict all those different things well at the same time. They tend to overfit quirks of one context. HCI, by being a distilled latent variable, generalizes better – an observation borne out by better cross-validated log-loss and R^2 across diverse tasks.

7.2.6 Ethics: minimize intrusion by maximizing efficiency

A stronger composite means we can often get the same accuracy with fewer sensors. For example, if HCI works well, maybe we find that three channels (say, self-report + HRV + pupil) are enough for everyday use, and we only need to add fMRI or EEG in special sub-studies. In contrast, a single weak measure might tempt us to throw the kitchen sink of invasive monitoring just to get decent accuracy – e.g. if “heart rate alone” were our metric, we might end up also demanding everyone wear an EEG cap and a cortisol patch to compensate for HR’s limitations, thereby increasing burden on participants. A composite that is efficient lets us be less intrusive overall. Ethically, that’s a win: we gather only what’s needed to reach a certain predictive confidence, and not more.

7.2.7 Robustness to adversarial and environmental drift

When conditions change (season, new phone firmware, shifts in incentive structures), any single channel can silently miscalibrate. For example, suppose a wearable’s algorithm update causes all HRV readings to skew high – if HRV alone were our metric, we’d falsely conclude everyone’s super chill now. In a composite, such discordance raises flags. We’d see the HRV channel’s loading a_k start to wander or its residuals blow up (because other channels like self-report and behavior wouldn’t agree with the sudden shift), or we’d find that using other channels to predict HRV now fails (see 7.1.7 on residual checks). This self-diagnosis is built-in: we know when channels disagree. With

a single meter, by contrast, you might have no clue anything went wrong until it's too late. The composite gives multiple points of comparison that make it much harder for a drift or sabotage in one channel to go unnoticed.

7.2.8 Multi-timescale truth requires multi-timescale sensors

The phenomena we care about span seconds to years. No single instrument covers that whole span. HCI is explicitly multi-timescale by incorporating fast, medium, and slow signals:

- Fast: Pupil dilation or EEG changes capture sub-second adjustments (e.g. a split-second of panic or relief).
- Medium: Self-reports and HRV capture minute-to-hour fluctuations (mood changes through a day, stress recovery over an afternoon).
- Slow: Behaviors and especially sleep/dream patterns reflect day-to-week adjustments (habits forming or breaking, a week of coping, REM rebound after a rough day).

The life ledger $L(t) = \int_0^t F(\tau) d\tau$ is intrinsically multi-timescale. If you tried to use, say, only EEG, you'd get a great millisecond view but lose the forest (what about yesterday's rebound?). Only surveys, and you might miss fleeting episodes that never got recorded. Composite HCI prevents aliasing – it won't mistake a slow drift for a burst or vice versa – because it explicitly models fast vs. slow components (see 7.4.2) and has inputs covering each relevant band.

7.2.9 Guardrails against circularity

Because HCI is built from multiple channels, we can do leave-one-modality-out tests to ensure we're not in a tautological loop. For example, we construct HCI minus EEG and see if that HCI still predicts EEG features; similarly HCI minus self-report should predict self-report scores. We have done these “channel holdout” validations (and will continue as data accumulate). We find that HCI-without-X predicts X to a useful degree for each channel X – meaning HCI isn't merely overfitting each channel's noise, it's capturing the common signal. A single meter obviously can't do this kind of check – it cannot predict itself from nothing. The composite's ability to “audit” each channel by omission is a unique guardrail against circular reasoning (e.g., claiming HCI predicts well-being when it was secretly just a re-label of self-report; we can show it's more than that).

7.2.10 Practical illustration (utility-matched repair vs. indulgence)

Scenario: You face two options that feel equally tempting in the moment. One is a repair (it addresses a lingering problem, high compensability), the other an indulgence (fleeting fun, low compensability). Both have the same immediate utility.

- Single self-report: You might just report “I feel torn, both sound good.” That’s ambiguous – we can’t tell which is truly better for you long-term from that alone.
- Composite HCl: Under the hood, multiple signals tilt one way. In this scenario, if LoF is correct, we’d expect: vmPFC (value area) shows a subtle extra glow for the reparative option (it “knows” this will feel better later); rIFG/STN (braking circuit) puts a quiet brake on the indulgence, especially as you contemplate a shorter future; HRV settles faster after choosing repair than it would after indulgence; next-day behavior (stickiness) is higher for the repair – you continue with that good trajectory rather than flaking. In fact, in tests, the composite called the tilt correctly: it indicated higher HCl for the reparative choice, aligning with which choice led to better next-day outcomes, whereas any single measure looked equivocal.

In short, the composite “saw” what the single measure couldn’t, and it predicted tomorrow’s behavior (feeling better and sticking to plan when the repair was chosen).

7.2.11 Where we go next:

Composite logic must be kept honest. 7.3 moves from theory to discipline—what we blind, how we preregister, and which adversarial checks stop us from smuggling in our hopes.

7.3 Keeping It Honest: Blinds and Preregistration

A measurement program is only as good as its anti-self-deception machinery. This section specifies the blinding, preregistration, and adversarial practices we require before we will treat any HCI or QS result as credible evidence. It's essentially a checklist to copy into protocols and IRB plans. If you're not doing these things, any positive result for LoF is provisional at best.

7.3.1 What must be blinded (and to whom)

Condition blinding:

- Participant-level: When we manipulate horizons or congestion (e.g. give a countdown timer or impose a waiting line), participants know something is happening but we mask the exact level. Through cover stories, we equalize expected value and arousal across conditions so they don't catch on that "this is the short-horizon high-stress condition." They might think it's just different game scenarios. For example, if we use a timer, they won't know if it's meant to create urgency or just a standard rule.
- Experimenter-level: The people running sessions see only labels like "Condition A/B" which don't reveal which is the short-horizon or high-congestion tier. The scheduling of conditions is randomized by someone else (or by computer) so the experimenter can't stack the deck or inadvertently signal anything.

Outcome blinding:

- Analyst-level: The data processing pipeline is locked and automated. It ingests de-identified data and computes all primary outcomes, producing results where group labels are hashed (nonsense strings). The analyst doesn't know which group is which until everything (modeling, QC) is finalized. Only then do we reveal "A was short horizon, B was long," for instance. This prevents p-hacking by peeking.
- Clinician/observer-level: In any field context (say we're in hospice or a clinic measuring "reconciliation" outcomes), the staff or raters who judge those outcomes are blind to our assignments and HCI values. For example, a nurse judging a patient's sense of closure shouldn't know whether we classified that patient as short-horizon or not.

Channel blinding:

- Our analysts are siloed by channel: the person extracting features from, say, brain scans has no access to self-report or behavior data, and vice versa. Only at the final fusion stage do we combine them. This reduces “leakage” where someone might unintentionally tweak one channel’s processing because they know another channel’s outcome. We basically treat each channel’s preprocessing as its own blinded pipeline until fusion.

7.3.2 Preregistration: what gets locked before first look

Before we collect or look at any data, we preregister the following components in detail:

- Hypotheses: We state the exact expected *directions* of primary effects. For example, we might write: “ $\gamma_{\Phi}(\text{vmPFC}) > 0$ ” (*the coefficient on compensability in vmPFC activity is positive*), “ $\gamma_{\{\Phi \times H^{-1}\}}(\text{rIFG/ACC}) > 0$ ” (*the interaction of compensability with short horizons increases braking activity in rIFG/ACC*), “stickiness asymmetry increases as $H \downarrow$ ” (*people stick more to high- Φ actions and drop low- Φ actions when horizons shrink*), “*dream inversion* $D_n \approx -a(H) \Delta L_{\text{day}}$,” with $a(H)$ growing as H shortens. Every primary pattern gets a prediction.
- Design and power: We justify our sample size with simulation-based power analysis (see 6.4.7 for our approach). Typically we target $> 90\%$ power to detect standardized effects around 0.10–0.20 (small but meaningful) under cross-validated analyses. This prevents underpowered studies that might only find false positives.
- Instruments and manipulations: We document the exact horizon manipulation (e.g. a countdown timer of 30s vs. 5m), congestion manipulation (e.g. presenting a queue of 2 people vs. 10 people ahead in line), along with credibility checks (questions ensuring participants noticed and believed these setups). We also define negative controls – “sham” versions that should have no effect (like a fake timer that isn’t actually limiting anything) – and set criteria for what constitutes a pass vs. fail for those controls.
- Backbone model: We list every nuisance covariate and base model term. For instance: utility, risk level, conflict, arousal indices, effort required, fatigue, learning effects, habit tendencies – all to be included as fixed effects or structured in the model. We specify which brain ROIs and time–frequency windows we’ll focus on, which preprocessing steps (e.g. filter settings, artifact criteria), and the rules for exclusions (like motion thresholds, compliance cutoffs). This way, nothing about the baseline model is decided after seeing data.

- QS terms: We lock in the exact features for LoF: which components of Φ (ReliefGain, RepairGain, Irreversibility, OptionFlex) will be used, how they'll be coded, how they'll interact with H^{-1} (time horizon inverse), and how shared resource penalties $\sum_r \lambda_{rt} \Delta_r(u)$ enter utility. No adding a surprise new term later because it looked good.
- Primary analyses: We enumerate them. For example: (a) Behavioral choice GLMs evaluated with blocked cross-validation (held-out days or persons) – we'll report LOO or WAIC or blocked ROC AUC/log-loss improvements. (b) Neuro analyses: fMRI ROI betas and EEG power contrasts tested with cluster-corrected stats at $\alpha=0.05$. (c) State-space model settings: what priors we will use for loadings (a_k) and process noise (σ_ξ) – essentially fixing the HCl model hyperparameters.
- Stopping rules: We explicitly state that we won't peek at results early. If, for safety reasons, we must have interim looks (e.g. in a patient study), we will use a-spending or Bayesian sequential criteria defined in advance. This ensures we can't bail out or declare victory early just because we saw a $p=0.049$ at midpoint.
- Spec-curve band: We set a small, predefined multiverse of reasonable analysis variations (say 8–12 variations: two plausible ways to score something, two plausible outlier criteria, etc.). We commit to reporting the full spread of outcomes across that multiverse. This way, if our result only appears under one very specific analysis and flips under a slightly different reasonable choice, readers will see that.

All these details are posted to an open registry (like OSF or AsPredicted) with a timestamp before data collection begins. If we ever need to change something (maybe a flaw is discovered midstream), amendments are logged and the original entry is not overwritten – full transparency.

7.3.3 Randomization, counterbalancing, and instruments that cannot cheat

- Blocked randomization: Within each participant, we randomize the order of horizon conditions and congestion loads so that each person experiences them in a counterbalanced way. Across participants, we might use Latin square designs to avoid any order or day-of-week effects.
- Instrument checks: We verify that our experimental instruments do what we think. Horizon “primes” (like time limits or life-event vignettes) must shorten perceived horizon H_t without jacking up arousal (we'll check pupil or SCL to confirm no unintended stress response). Congestion manipulations (like giving some participants extra “gatekeepers” to go through) must increase waits/denials (i.e. λ) without altering the inherent utility of actions. If our horizon manipulation turns

out to just be stressing people out, that's a failure of the instrument, not a test of LoF.

- Placebos: We include non-credible primes or sham devices that should, in theory, do nothing. For example, a “dummy countdown” that participants believe might matter but in reality doesn’t constrain anything. These must produce null effects. If our analysis pipeline detects effects under these sham conditions, we halt and fix the pipeline or instrument before proceeding. In other words, if we get a false positive on a placebo, our entire experimental setup is suspect.

7.3.4 The locked pipeline and code escrow

- Containerized analysis: We put our entire data processing and analysis code in a version-controlled container (e.g. a Docker image) with fixed package versions, fixed random seeds, and fully reproducible outputs. This environment is created *before* we see any outcomes. That way, when we hit “go,” everyone (including us) knows exactly what will happen with the raw data – no on-the-fly tweaking.
- Code escrow: At the moment of preregistration, we take a cryptographic hash of our analysis code and push it to a neutral third-party escrow or a public repository (with the hash documented). This proves later that we didn’t modify code after seeing the data (because any change would change the hash). When it’s time to unblind, we only change a configuration file to point to the now-labeled data – the code stays the same.
- Data provenance tracking: Every data file (raw, intermediate, final) gets an immutable fingerprint (hash). Our pipeline is set so that if someone tries to overwrite or alter a file (say, replacing a dataset after peeking), it triggers an alert or fails checks. This ensures no sneaky substitutions happen once the pipeline is locked.

7.3.5 Adversarial collaboration and red teams

- Rival PI role: For confirmatory studies, we often bring on board a respected external researcher (Principal Investigator) whose theoretical stance is opposed or alternative to LoF. This “rival PI” co-owns the nuisance models and contrasts. They have veto power on any questionable modeling choices (they help ensure we’re not stacking the deck). They also are invited to write a section in the report outlining how a lean rival model could explain the results if it can – essentially contributing an honest critique from the inside.
- Red-team drills: Before we unveil final results, we designate an internal team (or external experts) to act as attackers on the analysis. They’ll do things like randomly permuting condition labels, injecting bogus data, toggling

preprocessing choices – trying to break the result. If our claim only holds under very fragile conditions, these drills will expose it (e.g. if shuffling one participant’s data kills the effect, that’s worrisome). We only proceed to publish if the claim survives these stress tests intact.

- Blind prediction challenge: We go further to invite outside teams. We’ll release a portion of feature-extracted training data without labels, and challenge outsiders: “Using your own model or theory, predict the held-out test set’s outcomes.” After they submit predictions, we reveal the true labels and see how everyone did. If an outside team’s predictions (e.g. from a simpler model) are as good as HCI/LoF’s, that’s humbling information we must report. This is a way to ensure we haven’t overlooked a trivial explanation that others can spot.

7.3.6 Multiverse and robustness reporting

- Specification curve: We predetermined a small “multiverse” of analysis variations (different reasonable choices for data filtering, baseline subtraction, ROI definitions, HRV calculation, etc.). In publication, we present a specification curve plot – basically the distribution of effect sizes across all those choices. We emphasize the median effect and the interquartile range, not the best-case. If our claim only holds in the most favorable spec, that will be obvious. The reader sees the stability or brittleness of the result.
- Sensitivity to priors: For the Bayesian state–space HCI model, we report how much the results change under different prior assumptions. For example, if we loosen the prior on each channel’s loading a_k or on process noise σ_ξ , do the core inferences (like end-of-life drift going to zero) remain? This guards against us hand-tuning priors to get desired outcomes.
- Negative controls: All analyses of sham conditions and manipulations that should do nothing are reported in the main text. We don’t bury “oh by the way, our placebo had a weird effect” in an appendix. If the pipeline falsely found an effect where none should exist, we confront it transparently.

7.3.7 Data sharing and privacy

- Tiered sharing plan: We outline levels at which data can be shared without compromising privacy.
 - Tier 1: Fully synthetic datasets that mimic the real data’s statistical structure (for general teaching and code testing). These have no one’s actual data, just simulated points drawn from our model, so they pose no privacy issue.

- Tier 2: Anonymized, de-identified feature tables (e.g. per-trial or per-day summaries like “HRV = X, self-report = Y”) possibly with a bit of noise added for differential privacy. This allows other researchers to validate our analysis without ever seeing raw sensitive info.
 - Tier 3: Raw data (biosignals, etc.) available only under formal data use agreements and ethics board oversight. Basically, if someone has a really good reason, they can access raw data in a controlled environment.
- On-device privacy: As noted, text and audio remain on the participant’s device by default. Only their transformed features leave the device. We even share the embedding models so others can see how text was converted to numbers without ever reading the text.
- Right to withdraw enforcement: If a participant withdraws, we have systems to propagate deletions (e.g. removing their row from feature tables and excluding them from aggregated analysis). We maintain machine-readable logs of consent status to ensure no data is analyzed against a participant’s wishes.

7.3.8 When a result “counts”

Not every significant-looking result will be counted as a confirmed finding. We enforce that a result only counts if all of the following are true:

1. Backbone wins first: The baseline covariates (utility, conflict, arousal, etc.) explain what they are supposed to explain in the model. (E.g., if even basic effects like “harder tasks feel worse” isn’t observed, the whole experiment may be off.)
2. QS/LoF additions improve prediction: When we add the LoF terms (Φ features, $\Phi \times H^{-1}$ interactions, λ congestion terms) to the model, the held-out predictive performance improves appreciably. It’s not enough that they’re significant in-sample; they must make the model generalize better.
3. Sham controls show nulls and credibility checks pass: The negative control conditions yield no effect (as expected) and participants believed/took seriously the manipulations. If, say, a fake countdown oddly produced an effect, or if people didn’t buy the scenario, the result is disqualified.
4. Replication: We have either a second independent site showing the effect or a within-subject replication on held-out days for longitudinal studies. No result based on one batch of subjects or one week of data is considered conclusive.
5. Rival sufficiency check: No leaner model with equal or fewer degrees of freedom can match the result. If a simpler explanation (like a well-tuned RL model) can

predict everything HCI can, then our result doesn't prove we needed LoF – we would credit the rival.

6. Auditable and transparent: All code, data decisions, and even null results related to the finding are available for inspection. And we have reported the null results alongside positives. If any of these pieces are missing, the result isn't fully trustworthy yet.

Anything less than the above is labeled exploratory. We may share it as a hypothesis generator, but we won't call it evidence *for* the Law of Fairness until it passes all the gates.

7.3.9 Common failure patterns and the mandated response

We have also anticipated some common ways things could go wrong, along with pre-decided responses:

- Effect dies after adding arousal. *Pattern:* Our key effect (say, horizon shortening causing more reparative choices) disappears once we add an arousal covariate. *Response:* We classify that result as “arousal-substituted” – meaning what we found was probably just an arousal effect, not LoF. We then redesign the instrument to better isolate pure horizon changes (e.g. use a calmer horizon manipulation) and we do not claim a QS/LoF effect there.
- Single-site wonder. *Pattern:* We get a beautiful result in Lab A, but when Lab B tries, it's flat. *Response:* We freeze any grand claims and treat it as provisional. We must run a second site replication (or more) before publication. If it doesn't replicate, we'll publish Lab A as an anomaly or exploratory finding, not as law-confirming evidence.
- Spec-curve brittleness. *Pattern:* The spec curve shows that some reasonable data processing choices make the effect vanish or reverse (e.g. depending on whether we include a certain outlier or use one HRV metric vs. another, we get different outcomes). *Response:* We do not hide this; instead, we go back and *revise our preprocessing* until the effect is robust across the predetermined range – or we downgrade the finding to exploratory if it remains fickle.
- Rival sufficiency. *Pattern:* A rival model (like a predictive coding model with risk and fatigue) manages to explain the data as well as HCI does. *Response:* We explicitly state that in any report. We would then scale back our claim – acknowledging that LoF might not be necessary to explain that domain – or position the finding as “a tendency consistent with LoF, but also explicable by X.” In practice, if a leaner explanation fits, LoF doesn't get to claim a law-like status for that pattern.

7.3.10 One-page template (drop-in for your lab)

To encourage widespread rigorous testing, we've developed a template any lab can use (and we use ourselves) when designing an HCI/QS study. It's like a pre-experiment checklist:

Title: *HCI/QS Horizon-Scaling Study vX.Y* (versioned for any tweaks).

- Registry link: URL and hash of the preregistration document.
- Primary outcomes: Bullet list of the key coefficients or contrasts (e.g. “Short horizon $\times \Phi$ effect on choice ($\gamma = ?$), predicted positive”).
- Manipulations: Description of the horizon instrument, congestion instrument, and any sham conditions.
- Backbone model: List of covariates, brain regions of interest, time–freq windows, exclusion rules, etc., that form the baseline model.
- QS terms: Exactly which Φ features, interactions with H^{-1} , and any social penalty terms are included.
- Power: Summary of simulation-based power analysis and target sample size.
- Pipeline: Container or analysis code hash and location of the escrow (so anyone can verify the pipeline was pre-specified).
- Blinding: Who is blind to what (participants, experimenters, analysts) and how it's achieved.
- Stop rules: Any interim analysis plans, alpha spending rules or Bayesian thresholds for stopping, if applicable.
- Red team: Names of individuals or team responsible for adversarial testing, and scope of what they will try.
- Data sharing: Plan for Tier 1/2/3 data releases and timeline.
- Decision rule: Criteria for declaring success or failure (essentially the points from 7.3.8 above, tailored to the study).

This fits on one page and serves as a succinct contract of how the study will be conducted. We include it in our documentation so that anyone – including ourselves months later – can quickly see what was supposed to happen.

7.3.11 Where we go next:

Blinds and prereg set the floor; models set the ceiling. 7.4 opens the latent/state-space toolbox, explaining how to fuse channels into one trajectory without overfitting or wishful smoothing.

7.4 Research Notes: Latent (CFA/IRT) and State-Space

This section is a methods workbench for building the Hedonic Composite Index (HCI) as a latent variable with two complementary layers:

Formal model (compact form).

State equation: $S_{t+1} = S_t + u_t - r_t + \omega_t$, with $F_t = -W \cdot (S_{t+1} - S_t)/\Delta t$.

Here Δt is the sampling interval ($\Delta t > 0$) in the same time units used for ledger integration, and W is a fixed nonnegative weight (vector or scalar) that sets the sign and scale of F_t . Here u_t aggregates exogenous drive loads (e.g., pain, threat, deprivation), r_t endogenous relief/repair (sleep, social restoration, reappraisal), ω_t process noise. Measurement equation (multi-channel): $y_{k,t} = h_k(S_t, (S_{t+1} - S_t)/\Delta t, F_t) + \varepsilon_{k,t}$. Delta-HCI link: compute $HCI_t = \sum_k w_k \cdot \Delta z_{k,t}$ and set $F_t := HCI_t$ post-calibration. Read “ $:=$ ” as calibration: HCI_t is the reported estimator, and F_t is the latent state it estimates (with uncertainty carried through $p(F_t | \text{data})$). The latent layer (CFA/IRT) harmonizes indicators; the state-space layer enforces temporal coherence and yields $p(F_t | \text{data})$ needed for ledger integration.

Latent measurement layer (CFA/IRT): We harmonize multi-channel indicators to a single latent net-affect signal $F(t)$ with principled uncertainty.

A state-space layer – a time-series model that lets this latent factor evolve over time and fuses asynchronous data streams in real-time.

Use these notes to guide preregistration of models, choice of priors, checks for model identification, and drift diagnostics for HCI. (In more casual terms, this is the technical “how-to” for the math behind HCI.)

7.4.1 The measurement layer: CFA/IRT in brief

We treat each channel K (e.g. certain self-report items, HRV metrics, EEG components, fMRI ROI betas, behavioral frequencies, dream-affect scores) as an indicator of a common latent momentary affect F_t . There are two main cases:

- Linear CFA (continuous indicators): For approximately continuous or normally-distributed indicators (e.g. a z-scored HRV or a brain ROI beta), we use a standard factor model: $y_t^{(k)} = v_k + \lambda_k F_t + \Gamma_k Z_t + \varepsilon_t^{(k)}$, with $\varepsilon_t^{(k)} \sim \mathcal{N}(0, \Psi_k)$. Here $y_t^{(k)}$ is the observed value (e.g. HF-HRV at time t), v_k is an intercept, λ_k is the loading (how strongly channel k reflects the latent affect), Z_t are observed nuisance covariates for that channel (arousal, effort, etc.), and Ψ_k is the residual variance after accounting for F_t and covariates.

- Ordinal/threshold IRT (discrete indicators): For Likert-scale survey items or categorical tags (like dream themes), we use an IRT approach. For example, a graded response model: $\Pr(y_t^{(k)} \geq c | F_t) = \text{logit}^{-1}[a_k(F_t - b_{\{k\}c})]$, where a_k is the discrimination (analogous to a factor loading) and $b_{\{k\}c}$ are threshold parameters for item K across response category C. Intuitively, this says “the probability of rating at least C on item K is given by a logistic curve of F_t ” – so higher F_t leads to higher probabilities of higher responses, depending on thresholds $b_{\{k\}c}$.

Identification: We must set the scale of F_t so the model isn't arbitrary. Usually, we fix $\text{Var}(F_t) = 1$ (latent variance = 1) or fix one loading (e.g. $\lambda_{\text{EMA}} = 1$) to set the unit. Later, we will rescale F_t to Hedonic Composite Units (HCU) in 7.5, but at the measurement stage we work on a relative scale.

For multi-group scenarios (comparing people or labs), we do a multi-group CFA/IRT to ensure measurement invariance: we test configural invariance (same factor structure across groups), then impose metric invariance (λ_k equal across groups), and ideally scalar/threshold invariance (v_k or $b_{\{k\}c}$ equal). If scalar invariance fails, we can still proceed with partial invariance or alignment methods (basically allowing intercepts to differ but adjusting factor means accordingly) – but we then avoid comparing raw HCl levels across those groups. In short, we use the measurement layer to clean and standardize the signals first, before the time dynamics come into play.

Why do the measurement layer first? Because it prevents the time-series model from “trying to explain away” measurement artifacts as if they were real dynamics. By establishing a stable factor structure, we assign proper weights to each channel upfront. Then the state-space model can focus on real changes in affect, not on calibrating between HRV units and survey units on the fly.

7.4.2 The dynamics layer: state-space fusion

The latent affect is not static. We model its evolution and handle asynchrony and missingness with a state-space model.

Core model (linear-Gaussian):

State: $F_t = F_{t-1} + \alpha^T u_t + \eta_t, \eta_t \sim \mathcal{N}(0, \sigma_\eta^2)$

Observation: $\tilde{y}_t = \Lambda F_t + \Gamma Z_t + \varepsilon_t, \varepsilon_t \sim \mathcal{N}(0, \Psi)$

- u_t : exogenous inputs (analgesia, nociception, sleep stage, social repair events).

- \tilde{y}_t : vector of CFA/IRT-linearized indicators available at time t (some entries missing).
- Estimate with Kalman filter/smooth; use particle filters if we keep non-linear IRT likelihoods online.

Multi-timescale extension:

Capture fast/slow components:

$$[F_t^{\text{fast}} \ F_t^{\text{slow}}] = A [F_{t-1}^{\text{fast}} \ F_{t-1}^{\text{slow}}] + B u_t + \eta_t, \quad F_t = F_t^{\text{fast}} + F_t^{\text{slow}}.$$

- Constrain $F^{\{\text{fast}\}}$ with larger σ_η ; $F^{\{\text{slow}\}}$ with smaller σ_η .
- This prevents aliasing between momentary swings and baseline drift.

Non-stationary noise (sleep kernels): Let $\sigma_{\{\eta,t\}}^2$ shrink during SWS and expand during REM to reflect variance compression and counterweight processing:

$$\sigma_{\{\eta,t\}}^2 = \sigma_0^2 \exp(-\kappa_{\text{SWS}} 1_{\text{SWS}} + \kappa_{\text{REM}} 1_{\text{REM}}).$$

Horizon and congestion as exogenous states: From 6.4, include estimated horizon H_t and shadow prices λ_{rt} in Z_t . We do not let them drive F_t unless theory dictates (avoid endogeneity); they primarily adjust observation models and later choice models.

7.4.3 Hierarchical bayes: partial pooling and priors

We want subject-level flexibility with population-level discipline.

Loadings and thresholds: $\lambda_{\{k,i\}} \sim \mathcal{N}(\mu_{\{\lambda_k\}}, \tau_{\{\lambda_k\}}^2)$, $b_{\{kc,i\}} \sim \mathcal{N}(\mu_{\{b_{\{kc\}}\}}, \tau_{\{b_{\{kc\}}\}}^2)$.
Shrinks noisy individuals toward group means; reduces overfitting.

$$\sigma_{\{\eta,i\}} \sim \text{HalfCauchy}(0, 0.5), \psi_{\{k,i\}} \sim \text{HalfCauchy}(0, 0.5).$$

Input gains $\alpha \sim \mathcal{N}(0, 0.5^2)$ with weakly informative priors.

Why Bayes?

- Natural uncertainty propagation to ledgers and downstream tests.
- Straightforward multilevel invariance and cross-site pooling.
- Compatible with blocked cross-validation (PSIS-LOO).

7.4.4 From CFA/IRT to state-space: practical pipeline

1. Preprocess each channel (artifact rejection, z-scaling, nuisance subtraction).
2. Fit measurement model (CFA/IRT) on batched windows to learn loadings/thresholds; test invariance across groups.

3. Export linearized residualized indicators and posterior distributions over λ , b .
4. Fit state-space with those loadings as priors; run Kalman/particle smoothing to obtain F_t , \hat{F}_t and credible bands.
5. Anchor HCl units (see 7.5) using analgesia and nociception experiments.
6. Validate by channel holdouts (predict left-out channel from F_t , \hat{F}_t) and task/site holdouts.

After sentence:

Tip: keep a lightweight online filter (for live telemetry) and a full offline smoother (for publication-grade inference).

7.4.5 Identification pitfalls and fixes

- Drift vs. intercept creep. If v_k shifts slowly (device aging, season), the state-space can misread it as affect drift.
- Fix: include time-varying intercepts with random-walk priors per channel or add periodic covariates (to capture seasonality).
- Arousal masquerading as affect. Pupil/skin-conductance loadings may inflate if arousal is not included in Z_t .
- Fix: always include arousal covariates in observation models; test whether affect effects survive when conditioned on arousal.
- Over-tight priors. If $\tau_{\{\lambda_k\}}$ is too small, individuals cannot deviate; if too large, the model overfits noise.
- Fix: calibrate via simulation-based recovery; set τ to match realistic cross-person variance.
- Asynchrony bias. Channels sampled at different rates can dominate the latent estimate.
- Fix: subsample high-rate channels or inflate their ψ_k to equalize effective weight.

7.4.6 Model checking and multiverse robustness

- Posterior predictive checks (PPCs). Simulate indicators from the fitted model; compare simulated vs. real distributions, autocorrelations, and cross-correlations.
- Channel ablations. Refit the model without each channel in turn; measure degradation in predictive accuracy for held-out behavior and physiology.
- Spec curve. Vary plausible preprocessing choices (e.g. different HRV metrics, EEG baselines, motion thresholds) within a preregistered multiverse; report the median effect and interquartile range on key outcomes.

7.4.7 Out-of-sample evaluation

- Blocked CV: Use blocked cross-validation across days and subjects for one-step-ahead latent F_t forecasts and for downstream choice predictions.
- PSIS-LOO: For Bayesian fits, use PSIS-LOO; ensure $\hat{k} < 0.7$ (if diagnostics show $\hat{k} \geq 0.7$, refit with heavier-tailed distributions).
- Calibration curves: Plot predicted vs. observed indicators (and choices) to assess calibration.

7.4.8 Anchoring and unit rescaling (pointer to 7.5)

After estimation on an arbitrary latent scale, we rescale to Hedonic Composite Units (HCU):

- Map the median analgesia effect over a two-minute window to +1.0 HCU.
- Map the median cold-pressor rise (pain response) to -1.0 HCU.
- Apply a linear transform $F_t^* = a F_t + b$ to achieve these anchors (propagating uncertainty).

7.4.9 Minimal reproducible example (pseudocode)

Measurement layer

for each channel k:

preprocess $y_{kt} \rightarrow y_{k,t}, Z_t$

fit CFA/IRT to estimate $\{\lambda_k, v_k$ (or $b_{-k}c\}$, $\psi_k\}$ with group invariance

export posteriors for parameters and linearized indicators

State-space layer

initialize priors from CFA/IRT posteriors

for t in time:

predict $F_t | F_{t-1}, u_t$

update with available y_t (missing-safe)

smooth to obtain $\hat{F}(t)$ and $\text{Var}[\hat{F}(t)]$

Calibration

apply unit transform -> HCU

validate via channel holdouts, blocked CV, PPCs

7.4.10 When to stop and publish a null

- Loadings fail invariance across two independent cohorts and alignment cannot rescue comparisons.
- Channel holdouts show HCl does not predict left-out channels better than the backbone predictors alone.
- Downstream models show no improvement in predicting choices or physiology after adding F^t , \hat{F}_t and QS terms.
- Spec curve reveals effects are brittle to trivial preprocessing choices.

Nulls, honestly earned and fully reported, increase the credibility of subsequent positives.

Takeaway. CFA/IRT gives us a clean lens; state–space gives us temporal focus. Together they turn messy, asynchronous indicators into a transparent, auditable estimate of momentary affect that can be anchored, compared, and tested against the strongest rivals. This is the technical backbone that lets HCl—and any claim resting on it—stand up in hostile light.

7.4.11 Where we go next:

A composite isn't a unit. 7.5 defines HCU with behavioral and psychophysical anchors so that ledgers add up and thresholds mean something in practice.

7.5 Hedonic Composite Units (HCU)

To compare experiences across minutes, months, and people, HCl needs a unit. The Hedonic Composite Unit (HCU) is a calibrated increment of net affect—large enough to be behaviorally meaningful, small enough to resolve everyday swings. This section defines HCU, shows how we anchor it, and explains how to maintain the unit across sites, sensors, and time.

7.5.1 What an HCU is

Definition: One HCU is the standardized shift in latent affect that corresponds to a minimally clinically meaningful change in felt load/relief, verified by converging report, autonomics, and choice.

Use: HCU lets us express momentary affect F_t and the ledger $L(t) = \int_0^t F(\tau) d\tau$ on a shared, interpretable scale, enabling comparisons and pre-registered thresholds (e.g., “ $\geq +5$ HCU within 24 h after analgesia”).

7.5.2 Dual-point anchoring (Pain ↑, Relief ↓)

By convention, $+HCU$ corresponds to relief/comfort and $-HCU$ to added burden/pain. If the fitted latent scale is reversed (e.g., higher F_t indicates greater load), we flip the sign before anchoring so that positive HCU always denotes net relief.

We set the unit with two real-world anchors that are ethical, repeatable, and widely available:

- Cold-pressor nociception (Pain ↑).
- *Protocol:* forearm immersion in 0–2 °C water for a fixed duration (or to tolerance), with immediate HCl tracking.
- *Anchor:* the median latent increase during the final 30 s of immersion (across a reference cohort) is defined as -1.0 HCU.
- Rapid analgesia (Relief ↓).
- *Protocol:* fast-onset, low-risk analgesic (e.g. nitrous oxide for dental analgesia) with continuous HCl monitoring.
- *Anchor:* the median latent decrease over the peak 2-minute response is defined as $+1.0$ HCU. In this anchoring section, “increase/decrease” refers to the latent load state S; by the sign convention in 7.4, an S decrease corresponds to an upward shift in net affect F.

By fixing both directions, we reduce drift and avoid a unit that only fits one modality. The final transformation from raw latent F_t to HCU is linear:

$HCU(t) = a F_t + b$, $a = (1 - (-1)) / (\Delta F_{\text{analgesia}} - \Delta F_{\text{cold}})$, $b = -a (\Delta F_{\text{analgesia}} + \Delta F_{\text{cold}})$

/ 2. Uncertainty from both anchors propagates to the HCU band.

7.5.3 Behavioral criterion: the “just-meaningful” threshold

To ensure HCUs matter in life, we validate a criterion effect:

- A +1 HCU shift should increase the probability of choosing a reparative option over an equally valued indulgence by ~5–10 percentage points (pre-registered band), holding utility, conflict, and arousal constant. This is a task-family calibration target (a “just-meaningful” benchmark), not a universal constant across all cohorts and domains.
- Symmetrically, a –1 HCU shift should increase abort probability for low-compensability continuations under short horizons.

If calibration yields a smaller behavioral impact, we adjust anchor windows or cohort weighting until the criterion is met. Any adjustment to meet this criterion is treated as a versioned unit-definition update, done on separate calibration data (not the confirmatory test set) and documented before analysis.

7.5.4 Cross-site transport: reference panels and pooled priors

- Reference panels: Each site runs the anchor protocols on a local panel of 10–20 people each year.
- Hierarchical pooling: We maintain a pooled prior over anchor responses; site-specific transforms are regularized toward the global reference but allowed to deviate with documented reasons (equipment, altitude, climate).
- Drift alarms: If a site’s anchor medians deviate > 0.3 HCU from the global posterior for two consecutive quarters, recalibration is required before publishing HCl-based results.

7.5.5 Micro-anchors for daily life (non-lab)

Not everyone enters a lab or clinic, so we define micro-anchors that approximate ± 1 HCU in naturalistic contexts:

- SWS transition: The first consolidated slow-wave sleep bout of the night yields a canonical variance compression; its predicted reduction in HCl volatility over ~20 min provides a stability anchor.
- Breath-guided downshift: A standardized 5-minute paced-breathing session (e.g. 6 breaths/min) with an HRV increase above a person-specific threshold counts as a +0.3–0.5 HCU micro-anchor.

- Standardized irritant: A brief, ethically mild thermal or pressure stimulus applied at home (with consent) provides a $-0.3\text{--}0.5$ HCU micro-anchor.

Micro-anchors never replace lab anchors in publications, but they allow field recalibration between lab visits.

7.5.6 Ledger arithmetic with HCU

- Momentary flow: Report F_t in HCU at the chosen cadence (e.g., minute resolution). More precisely, $L(t)$ is the time-integral (area), so it carries HCU·time units.
- Accumulation: Ledger increments are time-weighted: $L(t_2) - L(t_1) = \int_{t_1}^{t_2} F(\tau) d\tau$ (HCU·time).
- Windows: Daily ledger L_{24h} and rolling weekly L_{7d} are standard summaries, each with uncertainty bands propagated from HCl.

7.5.7 Worked examples (intuition builders)

- Dental analgesia: Baseline HCl = -0.8 HCU (anticipatory load). Within 2 min of nitrous onset, HCl rises to $+0.4$ HCU; net change $+1.2$ HCU. The patient chooses to continue the procedure rather than abort.
- Cold pressor: Baseline HCl = $+0.2$ HCU. Final 30 s rises to -0.9 HCU; net change -1.1 HCU. Immediately after removal, HCl rebounds $+0.6$ HCU, demonstrating counterweight dynamics.
- Sleep counterweight: A difficult day ends at -2.3 HCU·day. Night REM density is high; the next morning's HCl baseline is $+0.6$ HCU higher than predicted by fatigue alone, consistent with dream inversion.

7.5.8 Handling medications, traits, and culture

- Medication adjustments: Known modulators (SSRIs, benzodiazepines, opioids) enter calibration as interaction terms on the anchors; we publish both as-treated and medication-adjusted HCU results.
- Trait stratification: For cohorts with chronic pain, depression, or anxiety, use stratified anchor priors; keep the HCU unit constant but allow baseline offsets in F_t (latent affect) per trait.
- Cultural invariance: When scalar invariance fails for self-report items (see Chapter 8), rely more on physiology and behavior in HCU; document the effective channel weights by group.

7.5.9 Uncertainty, error budgets, and reporting

Every HCU time series carries a credible band. We publish an error budget that decomposes variance into:

- Anchor error (between- and within-site)
- Measurement error (per channel)
- Model error (state-space process noise)
- Mapping error (site transform to global HCU)

Effect claims must exceed a minimal detectable change (MDC) computed from this budget (for example, ± 0.35 HCU over 10 min for a given setup).

7.5.10 Guardrails against “unit drift”

- Quarterly re-anchoring: Sites repeat anchor sessions each quarter; automated checks flag drift.
- Firmware/algorithm locks: Device firmware updates trigger mandatory post-update anchor runs before new data can enter pooled analyses.
- Back-compatibility: When anchor definitions are updated, we publish a conversion function so old HCU time series remain interpretable.

7.5.11 Ethical calibration

- Cold-pressor duration ceilings and analgesia dosing are IRB-governed; participants can stop any anchor at any time without penalty.
- Home micro-anchors are opt-in with clear “skip” options; no raw text/audio leaves the device—only derived features contribute to HCI/HCU.

7.5.12 What would make us change the unit

We will revise the HCU definition if any of the following replicate:

- The ± 1 HCU anchors fail to produce the pre-registered behavioral criterion change.
- A different pair of anchors (e.g. capsaicin pain and transcutaneous vagal stimulation) yields superior cross-site stability and predictive performance.
- The current anchors systematically bias specific populations (e.g. due to skin or autonomic differences) even after nuisance modeling.

HCU turns the latent HCI into a shared currency backed by physical anchors, behavioral meaning, and transparent uncertainty. It makes ledgers comparable across people and time without pretending feelings are simple. With HCU in place, we can talk about how much better or worse a day felt in a way that travels from clinic to lab to daily life—and stands up in peer review.

7.5.13 Where we go next:

Units are only credible if they can fail. 7.6 lists the conditions under which HCI should be downgraded or rebuilt—and what to publish when the result is a principled null.

7.6 Fail Conditions for HCl

A good instrument earns its keep by making it easy to lose on honest tests. This section specifies clear, preregistered fail conditions for the Hedonic Composite Index (HCl) and what we must do when any are met. Think of these as circuit breakers: when tripped, we pause claims, diagnose, and either repair or retire the instrument.

7.6.1 Measurement validity failures

- M1 — Arousal substitution. After adding preregistered arousal covariates (pupil, SCL, HR, respiration) to the observation model, the incremental predictive value of F^t , \hat{F}_t for behavior and neural outcomes drops to zero (held-out ROC/log-loss improvement ≤ 0.01).
- Action: Rebuild instruments to shift horizons without arousal influence; if three preregistered replications remain null, label HCl an arousal meter for that paradigm and withhold LoF/QS claims in that context.
- M2 — Single-channel collapse. Channel holdouts show that HCl *without* channel X cannot predict X, *and* that HCl *with only* channel X predicts all other channels as well as the full HCl.
- Action: Remove non-predictive channels; if a single channel explains the multiverse, rename the construct (e.g. “enhanced HRV index”) and cease cross-channel generalizations.
- M3 — Non-identification. CFA/IRT loadings fail to converge or flip signs across sites; model fit indices remain poor (e.g. RMSEA $> .10$, CFI $< .85$) under multiple plausible specifications.
- Action: Treat the latent structure as unidentified; revert to channel-specific analyses until a stable factor emerges.

7.6.2 Invariance and transport failures

- I1 — Scalar invariance failure. Self-report items show non-invariance of thresholds across languages/cultures that cannot be rescued by alignment optimization; physiology/behavior loadings also drift beyond tolerances.
Action: Publish site-specific HCl variants; do not pool ledgers or report cross-group means.
- I2 — Anchor drift. Yearly anchor panels at a site deviate from the pooled anchor distribution by > 0.3 HCU in two consecutive years (or by > 0.5 HCU in a single year).
Action: Freeze new data intake; recalibrate devices and protocols; reprocess affected windows with updated transforms; disclose the drift in manuscripts.

- I3 — Device/firmware sensitivity. A minor device or firmware update changes channel loadings or residuals such that held-out predictive performance drops > 20% relative to the previous version.
Action: Version-lock devices; re-run anchors after updates; segregate pre- vs. post-update data in analyses.

7.6.3 Predictive and concurrent validity failures

- P1 — No out-of-sample gain. Adding F_t , \hat{F}_t to preregistered backbone models (utility, conflict, arousal, effort, fatigue, learning, habit) yields no improvement in blocked out-of-sample prediction of:
 - Choice among utility-matched options;
 - Stickiness vs. abort rates on multi-step policies;
 - Autonomic settling (recovery dynamics);
 - vmPFC/rIFG/ACC/insula activity
- Action: Downgrade HCI to *exploratory* status; do not use it to test LoF/QS until a redesigned index passes P1.
- P2 — Horizon-insensitivity. Across tasks and telemetry, coefficients tied to $\Phi \times H^{-1}$ remain indistinguishable from zero after power checks and manipulation credibility tests (see 7.3).
- Action: Count this as a strike against QS/LoF for horizon scaling; report the null result prominently; revisit instruments (see 6.6).
- P3 — Dream inversion null. With PSG-confirmed REM sleep and strong recall, dream-affect integrals D_h do not invert the day's ledger change ΔL_{day} nor scale with horizon in three independent cohorts.
- Action: Remove dream data from HCI for that population; revise the “low-cost counterweight” claim to a mere tendency or withdraw it entirely.

7.6.4 Robustness and multiverse failures

- R1 — Spec-curve brittleness. Effect sizes for primary outcomes swing across the preregistered multiverse such that the median crosses zero or the interquartile range spans both positive and negative values.
Action: Declare the result unstable; tighten preprocessing choices; if still unresolved, downgrade the claim to exploratory.
- R2 — Rival sufficiency. A lean rival model (e.g. one based on predictive coding with risk and fatigue, or a resource-rational RL with queue costs) matches or surpasses HCI-based models on held-out prediction with fewer degrees of freedom.

- Action: Prefer the rival model; narrow the LoF/QS scope; use HCI only where it adds unique predictive value.
- R3 — Cross-site inconsistency. Primary effects appear only in one site or vanish under minor task changes, despite matched power and pipelines.
- Action: Suspend general claims; initiate adversarial collaborations; if inconsistency persists, restrict claims to site-specific contexts.

7.6.5 Ethical and feasibility failures

- E1 — Consent/Privacy non-compliance. Tiered consent, on-device data minimization, or withdrawal mechanisms fail audit.
- Action: Halt data collection; purge non-compliant data; resume only after an independent review certifies the fix.
- E2 — Burden exceeds benefit. Participant burden to maintain the channels required for reliable HCI exceeds IRB limits or causes > 30% attrition before endpoint.
Action: Drop the burdensome channels; revert to the ethical-minimum set (Self-Report + HRV + Pupil) and accept the wider uncertainty.

7.6.6 Decision rules: downgrade vs. abandon

- Downgrade to exploratory if any two of M1, P1, R1, R3 occur in preregistered studies with adequate power.
- Remove specific channels if they repeatedly trigger M2, I3, or P3 while others remain healthy.
- Abandon HCI for LoF/QS testing in a domain if three or more of P1, P2, R2, I1 replicate across labs. Publish a negative “failure” paper and redirect effort to alternative formulations.

7.6.7 Troubleshooting flow (condensed)

- Verify instruments: Did horizon/congestion manipulations produce the expected change without increasing arousal? If not, fix the manipulation.
- Audit nuisances: Are backbone covariates complete and preregistered? If not, add the missing ones.
- Re-anchor: Run the anchors; check for drift; re-scale to HCU (with propagated uncertainty).
- Channel ablation: Identify culprit channels; drop or down-weight overly noisy channels.

- Rival bake-off: If a lean rival model wins cleanly, prefer it and rewrite claims accordingly.

7.6.8 Reporting obligations

- Transparent nulls: Archive and submit null results with the same prominence as positives; include sham conditions and manipulation checks in the main text.
- Error budgets: Publish HCI/HCU uncertainty decompositions (error budgets) for each study.
- Versioning: State the HCI/HCU version, anchor dates, device firmware version, and pipeline hash in every manuscript.

7.6.9 Where we go next:

With HCI defined and guarded, Chapter 8 tests whether the same measurement truly travels across people and places. If the scale drifts across groups, we stop comparing and fix the meter before we make claims.

Chapter 8 — “Same Scale” Across People and Places

Fairness isn’t just about where you stand today — it’s about the road you’ve traveled. Imagine two people with very different journeys: one suffered through a painful childhood but is comfortable in old age, while another cruised through youth but faces chronic pain later on. If we only look at them *right now*, we might miss the bigger picture of who has endured more suffering overall. Chapter 8 is all about accounting for fairness over time. If HCl is like a speedometer (telling us the current comfort level for a person), then this chapter introduces the odometer and the operational tests that govern our fairness checks. We’ll talk about units (the basic “atom” of relief we measure), ledgers (the running total of those units each person has accumulated), and neutrality gates (checkpoints where each ledger is evaluated against pre-specified equivalence bounds $\pm K$ HCU at the death of mind).

In plain terms: Imagine every act of relief or comfort given to a person is like a coin dropped into their personal piggy bank. Hedonic Composite Units (HCU) are the standardized unit of ledger accounting—a calibrated measure of net affect. An HCU is defined via anchored HCl \times time; for example, a sustained increase in HCl mapped by the ± 1 HCU anchors (Chapter 7) integrates to a corresponding HCU amount. Throughout this chapter, HCl(t) is treated as a momentary rate expressed in HCU per unit time, so integrating HCl over time yields totals in HCU. The key is that we establish a common unit so we can add things up meaningfully. Once we have units, we keep a ledger for each person: a running tally of net HCU (positive and negative) accumulated over time. Mathematically, if HCl(t) estimates net momentary affect at time t, then the measured ledger is:

$\hat{L}(t) = \int_0^t HCl(\tau) d\tau$ is the area under the HCl curve—the cumulative net affect experienced up to time t. Here $\hat{L}(t)$ denotes the estimated ledger from measured HCl; when needed, we reserve L(t) for the (unobserved) true ledger. This ledger is the unit of assessment for LoF: the cumulative net affect per unified stream, not just moment-to-moment swings. Under LoF, each unified stream’s ledger is predicted to converge toward neutrality at the terminal checkpoint (within $\pm K$ HCU).

Because evaluation occurs at the terminal checkpoint, we use neutrality gates to test convergence. Think of neutrality gates as checkpoints or finish lines where fairness gets evaluated with a fine-toothed comb. The decisive gate is at the death of mind. LoF predicts that by the death of mind, each person’s ledger $\hat{L}(T)$ will be within a pre-specified $\pm K$ HCU band of zero for that unified stream. In practice, this means each ledger should be statistically equivalent to neutral within the preregistered $\pm K$ HCU band at the final gate. We preregister quantitative equivalence bounds for terminal tests (see Chapter 11).

We illustrate with bounds such as mean neutrality within $\pm 0.15 z$, trend within $\pm 0.05 z/\text{day}$, and $\geq 20\%$ variance reduction versus baseline. Here z refers to a standardized HCl scale used for near-terminal diagnostics; the terminal neutrality claim remains $\bar{L}(T)$ within the preregistered $\pm K$ HCU band. In plain terms, by life's end everyone's comfort level should have converged tightly. No one should have a lingering deficit of relief. These numbers put concrete bounds on what "fair enough" means. If those bounds are not met as the terminal checkpoint approaches, LoF would predict intensified compensatory pressures; our role is to measure, report, and test those dynamics. (Earlier neutrality gates might be set at interim checkpoints – say, every few years or at major life milestones – to make gradual adjustments, but the end-of-life gate is the ultimate backstop where LoF demands near-neutrality of outcomes.)

What compensatory patterns would we expect as the system approaches balance? It could mean giving extra attention, resources, or care units to those whose ledgers are falling behind. On the flip side, it might adjust the allocation of care for those far *ahead* (not to withdraw comfort they would get, but to ensure that boosting others does not come at the expense of anyone's dignity or baseline comfort). We'll also discuss how the system measures and models the distribution of these units of care. For instance, if we look at how many "comfort interventions" happen per day for each person, we treat that as count data. Throughout this chapter, λ used as a Poisson rate parameter (counts per unit time) is distinct from $\lambda_r(t)$, the shared-resource shadow price introduced elsewhere; the symbols overlap by convention, but the quantities do not. We might start by modeling such counts with a Poisson process—i.e., an average rate (r) of, say, comfort actions per day for a person. We'll check this assumption by looking at dispersion: if the data's variance is much higher than its mean (a telltale sign of over-dispersion), then a straight Poisson model isn't a good fit. A simple rule of thumb: if the variance-to-mean ratio exceeds about 1.2, we consider it over-dispersed. This 1.2 cutoff is a heuristic; we also check dispersion with formal diagnostics and residual fit before switching model families. In our analyses, when we detected over-dispersion (for example, some individuals had highly erratic bursts of many interventions on some days and very few on others), we switched to a more flexible Negative Binomial model. The Negative Binomial can handle the extra variability by introducing its own dispersion parameter. (We note in the methodology that we use a log link function for these count models, which is standard practice.) In short, we have a statistically sound way to estimate how many "units" of relief people get, to identify if some are getting significantly more or less than others, and to adjust for any anomalies. For example, if one person needs far more interventions than others, that flags a special case the system needs to examine and address.

Throughout this chapter, we never lose sight of the ethical dimension. We're effectively putting numbers to compassion – accounting for kindness in a quantitative way – but we remember that people are not numbers and relief is not just a line item. So even as we treat comfort units rigorously, we uphold the golden rule stated earlier: *comfort and dignity override data collection*. If someone is at the end-of-life neutrality gate and in their final days, the system's focus is on providing relief, not on ticking metric boxes. The metrics serve the person, not vice versa. In fact, LoF's design aligns naturally with good care: by meeting those neutrality gate criteria, we inherently ensure each person's last days are as comfortable as modern care can make them (because the only way to equalize comfort by the end is to maximize comfort for anyone who's behind). This is a design intent rather than a guarantee: clinical feasibility and patient preferences can limit what is achievable even when neutrality targets are pursued. We'll discuss a real-world implication: LoF effectively encodes a principle similar to hospice philosophy, but with a quantitative twist – everyone should get to finish life's race of comfort at nearly the same place, and that place should be as *high* as possible for all.

What you'll get from this Chapter:

- Defining the unit of fairness: Understand what we choose as the basic unit of relief/comfort in our system. We explain it in intuitive terms (for example, a unit could correspond to a standard amount of pain reduction or a standardized “dose” of comfort), ensuring everyone knows exactly what we’re counting and comparing.
- The fairness ledger: Learn how each individual’s experience is tracked over time on a ledger. We introduce the simple integration formula $\bar{L}(t) = \int_0^t HCl(\tau) d\tau$ and translate it into plain English – basically adding up all the bits of comfort a person experiences to maintain a running total.
- Neutrality gates explained: A clear explanation of what neutrality gates are and why they matter. We detail the end-of-life neutrality gate with the exact fairness criteria: mean difference within $\pm 0.15 z$, slope difference within $\pm 0.05 z/day$, variance ratio ≤ 0.80 versus baseline. You’ll see why these numbers were chosen and how they translate into actionable rules. (For instance, if Person A’s ledger mean is 0.3 z below Person B’s as final days approach, that’s outside the ± 0.15 range – the system must intervene.) We’ll also mention any interim gates, and how they help gradually steer everyone toward fairness well before the final check.
- Modeling interventions as data: An overview of how we analyze the distribution of relief units or interventions. This includes our use of Poisson models for expected counts and how we check for over-dispersion. If the data show more variability than Poisson allows (variance $> 1.2 \times$ mean), we explain our switch to a Negative

Binomial model (with a log link) to better fit the data. This ensures our statistical models of “fairness units” are accurate and aren’t underestimating uncertainty or individual variability.

- Real-world implications and ethics: A discussion of what all these measurements mean for actual care. By quantifying units and ledgers, caregivers and system designers can spot who might be underserved or falling behind. We also emphasize how LoF’s quantitative approach complements humane care – for example, hitting the neutrality gate criteria inherently means everyone receives high-comfort care at life’s end, aligning with palliative best practices. And if there’s ever a conflict between “hitting the numbers” and a patient’s immediate dignity or comfort, LoF dictates we choose the latter. In other words, the metrics are tools, not masters.

Subsections in this Chapter:

- **8.1 The Invariance Problem** - Why cross-group comparability matters and how we test it: the HCI/HCU measurement model must behave the same way across languages, devices, ages, and sites, or cross-group claims collapse into artifacts rather than true experience differences.
- **8.2 Culture and Age Effects** - How norms, language, development, and aging shape *measurement* (not necessarily feelings) and the modeling toolkit (item design, response-style adjustments, bilingual bridges, partial invariance, greater weight on non-verbal channels) we use to keep comparisons honest.
- **8.3 Universal Anchors (Pain, Chills, Social Exclusion)** - Shared yardsticks that let us map “+1 HCI” in one group to “+1 HCI” in another: controlled nociceptive pain, aesthetic chills, and social exclusion, chosen for universality, parametric control, multichannel footprints, and ethics.

Where we go next:

We start with the foundation: in 8.1 we tackle the invariance problem directly—what it takes, in practice, for a “+1” in one person or place to mean the same felt change in another. That section sets the rules for when cross-group comparisons are legitimate and when we must keep claims local. From there, the rest of the chapter builds outward: culture and age, universal anchors, formal tests, a calibration ladder, and finally how all uncertainty is carried into the ledger.

8.1 The Invariance Problem

If two people truly feel the same latent affect at the same moment, our instrument should return the same HCl value – regardless of their language, culture, age, device, or environment. Measurement invariance is the name for this requirement. Without invariance, any cross-group comparison of HCl or HCU can collapse into an artifact of wording, physiology, or hardware differences rather than a real difference in experience.

In practice, invariance means that the HCl's measurement model behaves the same way across different groups or conditions. Formally, let F_t denote the person's *true* momentary affect (the latent factor we care about), and let $y_{\{g,t\}^{\{k\}}}$ be an observed indicator (like a specific survey item, sensor reading, or behavioral measure) for group g (which could be a language group, a site, a device type, an age bracket, etc.) and channel k . A simple linear factor model (CFA) and an ordinal IRT formulation make this concrete:

- Linear CFA model: Each continuous indicator is expressed as:
$$y_{\{g,t\}^{\{k\}}} = v_{\{k,g\}} + \lambda_{\{k,g\}} F_{\{g,t\}} + \Gamma_{\{k,g\}} Z_{\{g,t\}} + \varepsilon_{\{g,t\}^{\{k\}}},$$
 where $v_{\{k,g\}}$ is the intercept in group g , $\lambda_{\{k,g\}}$ is the factor loading (sensitivity of that channel to F) in group g , $\Gamma_{\{k,g\}} Z_{\{g,t\}}$ represents any added effects of nuisance covariates Z (like arousal or context) in group g , and $\varepsilon_{\{g,t\}^{\{k\}}}$ is measurement error.
- Ordinal IRT model: For an ordinal or binary indicator (e.g. a Likert item), we model the *probability* of a response at or above category C via a graded response model:
$$P(y_{\{g,t\}^{\{k\}}} \geq c | F_{\{g,t\}}) = \text{logit}^{-1}[a_{\{k,g\}}(F_{\{g,t\}} - b_{\{k,gc\}})],$$
 where $a_{\{k,g\}}$ is the slope (analogous to a loading) and $b_{\{k,gc\}}$ are the threshold parameters for category C in group g .

Using such models, we define levels of invariance:

- Configural invariance: Each group has the *same pattern* of factors and loadings (i.e. the same set of indicators measures the factor in each group), but the numeric values of loadings (λ) or intercepts (v) can differ by group. Configural invariance establishes that the basic *structure* of the construct is consistent. (If configural fails, it means people might be interpreting items so differently that the factor doesn't even exist in the same form across groups.)
- Metric invariance: The factor loadings are equal across groups ($\lambda_{\{k,g\}} = \lambda_k$ for all k). Under metric invariance, a one-unit change in the latent factor F produces the same change in an indicator's expected value in every group. This allows us to compare relationships (correlations or regression slopes) involving F across groups, because the *unit* of F is now consistent. However, group means are not comparable yet.

- Scalar invariance: Both loadings and intercepts (for CFA) or thresholds (for IRT) are equal across groups ($v_{\{k, g\}} = v_k$ and $b_{\{k, gc\}} = b_{\{kc\}}$). Scalar invariance means the zero-point of the factor is the same and observed scores have the same meaning across groups. This is required to compare latent factor *means* or absolute HCl/HCU values between groups. (A stricter level sometimes noted is strict invariance, where residual variances $\psi_{\{k, g\}}$ are also equal across groups, but this is rarely necessary for our purposes.)

For HCl, we require at least metric invariance for comparing patterns or correlates of affect across groups, and scalar invariance (at least partial) for comparing levels (e.g. average HCU or ledger totals) across groups. If only metric holds but scalar fails, we might trust comparisons of *changes* or *within-person effects* but not raw scores across groups.

8.1.1 Why invariance can fail (and how we detect it)

Common failure sources: There are many reasons HCl might not operate identically across groups:

- Language and idiom differences: The connotations of words differ. For instance, the word “anxious” might imply *eager anticipation* in one language versus *nervous distress* in another. Terms like “heavy heart” or “feeling open” might not translate cleanly.
- Response style biases: Some cultures or individuals tend to agree with statements (acquiescence bias) or use the extremes of scales more or less often. Others have different reference frames – e.g. one group might rate a 5/10 for pain as what another calls 7/10, based on different internal baselines.
- Physiological differences: Skin conductance readings can vary with climate or skin properties; baseline heart rate variability might differ by fitness or genetic factors. What constitutes a “normal” sensor reading in one group might be systematically higher or lower in another.
- Device or firmware differences: If different groups are measured with different hardware versions or sensor settings, those can introduce systematic shifts (one device might read higher GSR values than another, etc.).
- Contextual factors: Differences in fasting, caffeine use, sleep patterns, or norms about emotional display can all affect measurements. For example, if one site collects data in the morning and another at night, circadian effects could shift the signals. Medication usage patterns (e.g. a country where beta-blockers are common) could alter physiology in one group more than another.

Detection toolkit: We employ several analyses to check invariance and pinpoint problems:

- Multi-group CFA/IRT: We fit the measurement model simultaneously to all groups and impose increasingly stricter constraints (configural → metric → scalar invariance). We then check model fit indices and *modification indices* (suggested model adjustments) to see where misfit occurs. A drop in fit when going from configural to metric, or metric to scalar, indicates non-invariance in some loadings or thresholds.
- DIF (Differential Item Functioning) tests: For each item or channel, test whether groups with the same underlying F have different item responses. For instance, using item-response theory logistic regression, test if group membership has a significant effect on an item after controlling for F. Significant DIF means that item behaves differently (e.g. consistently higher or lower in one group at the same F).
- MIMIC models: Include group as a covariate directly predicting each indicator (while controlling for F) in a single-group model. A significant direct effect of group on an item suggests bias (akin to uniform DIF).
- Alignment optimization: When full scalar invariance is untenable, use statistical alignment methods (as in Mplus or other software) which allow small differences in thresholds and intercepts and attempt to find a scoring that maximizes agreement across groups. Alignment can identify which specific items are causing non-invariance and provide approximate comparability if the non-invariance is minor and spread out.
- Longitudinal invariance: We also check invariance over *time* within the same individuals (e.g. before vs. after a device swap, or across seasons) to ensure the instrument is stable over time in one person. If it isn't invariant over time in the same people, it certainly won't be across different people.
- Device linking designs: If invariance issues stem from device differences, we use methods like NEAT (Nonequivalent groups with Anchor Tests) or common-participant designs to adjust scales. For example, have a subset of participants wear both the old and new device and use their data to derive a transform linking the two measurement scales (see 8.4.6).

In short, we throw a battery of tests at the data before assuming Person A's "HCl = 5" is the same as Person B's "HCl = 5." If the tests show problems, we either fix them (e.g. adjust the model, drop or replace biased items, recalibrate devices) or we refuse to make cross-group comparisons with that data.

8.1.2 How we achieve “same scale” in practice

Ensuring invariance is both about designing HCl carefully and statistically verifying it:

- Step 1 – Build for invariance from the start. We craft HCl to be as language- and culture-agnostic as possible. That means using short, concrete self-report items with minimal slang or idiom (e.g. “My body feels tense right now” rather than abstract terms). We incorporate multiple indicator types (self-report, autonomic signals, brain signals, behaviors, even dream content) so no single channel (which might have group-specific quirks) carries the whole meaning. We also include nuisance covariates Z_t (like arousal level, effort exerted, ambient temperature, medication status) in the measurement model to soak up known differences – so groups that differ in, say, average arousal won’t appear different on F if it’s just an arousal effect.
- Step 2 – Physically anchor the unit. We bind the abstract scale of HCl to physical reference points that are as universal as possible. Specifically, we set a reference such that a specific pain stimulus corresponds to a -1.0 HCU drop and its relief corresponds to a +1.0 HCU rise (details in 8.3 and 7.5). These ± 1 HCU anchors are based on nociceptive pain (like the cold pressor test) and rapid pain relief (analgesic administration). Because physical stimuli like extreme cold or relief from pain have similar effects on humans regardless of language, these anchors help “ground” the scale across groups. In short, we use a yardstick from nature – pain is pain – rather than just words.
- Step 3 – Fit group-specific models then impose equality. In analysis, we first let each group have its own parameters ($v_{\{k, g\}}$, $\lambda_{\{k, g\}}$, etc.) when estimating the measurement model, to see how they differ. Then we progressively impose constraints (e.g. set all $\lambda_{\{k, g\}}$ equal) to test metric invariance. We always keep aside a holdout sample or use cross-validation: invariance isn’t just about fit in the estimation sample, but also that the model predicts new data equally well for each group.
- Step 4 – Repair instead of forcing, if needed. If full invariance fails, we try partial invariance: allow certain intercepts or thresholds that are clearly biased to differ for that group, while keeping loadings equal. This often salvages comparability – e.g. maybe one particular self-report item doesn’t translate well, so let that item have a different baseline in that culture, but everything else is the same. We also use alignment when many small differences exist: this technique finds an optimal compromise of parameters across groups without assuming exact equality, effectively down-weighting aberrant items. The philosophy is not to force invariance by fiat, but to adjust and be transparent when differences exist. If a

stable solution with partial invariance or alignment can be achieved (with minimal impact on factor means), we accept that and document it.

- Step 5 – Link devices and sites through overlap. When different data-collection setups are involved (multiple device types, labs, or versions), we use linking studies. For devices: run a subset of participants on both old and new devices (common-person design) or have some common calibration tasks across devices (common-item) and derive a linear transform so that readings from Device B can be converted into the units of Device A. For language/culture: include a small bilingual sample or use “vignette” anchors that can serve as common items across translations. Essentially, we create a bridge so that all sites can express their HCU in a common unit (with known uncertainty).
- Step 6 – Publish diagnostics and guardrails. We never just assume invariance; we report it. In each analysis or paper, we include appendices showing the fit indices for invariance tests, which items (if any) were freed for partial invariance, and how much difference that made. We also report the effective weights each channel got in each group (e.g. maybe heart rate counted a bit more in one group if GSR was noisy there). This transparency ensures that if someone sees a cross-cultural comparison, they can judge whether the scale was truly “same” or if caveats apply. Moreover, we establish rules for ourselves: if we cannot achieve at least partial scalar invariance without contorting the model, we will not compare means across those groups.

Finally, if scalar invariance simply cannot be achieved (even after partial adjustments or alignment) without wrecking the model, then we do not compare absolute HCU levels between those groups. We might restrict ourselves to comparing *within each group* changes or use a different analysis altogether. The integrity of LoF claims rests on not mixing apples and oranges. By securing invariance via design and testing, we aim to ensure one person’s HCU is another person’s HCU in meaning. If that fails, we openly narrow the scope.

8.1.3 Anticipating special cases

Some specific scenarios require extra care to maintain invariance:

- Children and adolescents: Developmental differences mean younger individuals may interpret questions differently or have different physiological baselines. We use developmentally tailored items (more concrete, simpler language, focusing on bodily feelings rather than abstract emotion words). We rely relatively less on self-report (since young children might lack vocabulary or introspection) and more on behavior and physiology (e.g. activity level, facial expressions) for their

HCI. We explicitly test for age-by-item DIF, since certain items might be harder for kids. Essentially, we treat age groups almost like different “cultures” in invariance testing.

- Older adults / neurodegenerative conditions: Aging can bring sensory loss (e.g. reduced pain sensitivity or hearing), different autonomic baselines, and many are on medications that affect mood and physiology. In these groups, we may down-weight channels that become unreliable (if someone has neuropathy, GSR may not react to pain the same way) and incorporate caregiver reports or passive monitoring when self-report is unreliable. We check longitudinally within-person to see if measurement properties shift as someone ages or as a neurodegenerative disease progresses – ensuring we don’t confound true affect changes with instrument drift.
- High-altitude or tropical field sites: Environmental extremes can systematically skew readings – e.g. at high altitude or in very hot/humid climates, baseline heart rate or skin conductance may differ. We include environment variables (temperature, altitude) as covariates and often rely on device auto-calibration (like temperature-compensated GSR readings). We also often re-run anchor tasks in each site to calibrate for any environmental effect (for instance, maybe at a high-altitude site the cold pressor yields a slightly different average response; we account for that in the HCU transform).
- Cross-language rollouts: Translating survey items and instructions can itself introduce non-invariance. We do careful back-translation (translating back into the original language to check fidelity) and include standardized vignettes or anchor descriptions that accompany numeric scales so that “5/10” means the same reference scenario in each language. Additionally, having a bilingual bridge sample – participants fluent in both languages rating in both – allows us to equate the scales between languages directly.

We design HCI/HCU with invariance in mind and verify it with data. It’s an ongoing, active process: as we expand to new populations or devices, invariance must be re-confirmed. This is not just bookkeeping – it is fairness in measurement. If our meter under-reads the suffering of one group or over-reads the pleasure of another, we’d be introducing injustice into the science. So whenever in doubt, we choose to widen uncertainty, lower our claims, or keep comparisons local rather than pretend the scale is universal when it isn’t.

8.1.4 Decision rules for pooling vs. stratifying

After running these tests, how do we decide if we can aggregate data or need to keep groups separate? We have clear decision rules:

- We allow pooling data across groups (i.e. treat them on one common HCU scale) only if all of the following hold:
 - Configural and metric invariance.
 - Scalar or threshold invariance holds fully or sufficiently ($\leq 25\%$ items with manageable thresholds).
 - Anchor stimuli are equivalent (e.g., mean HCU drop for cold pressor within 0.3 HCU of the pooled reference).
 - Devices/firmware are identical or appropriately linked.

If any fail, we do not pool. We stratify analyses or use multi-group models and explicitly state that cross-group comparisons are not on a shared scale. The take-home: “same scale” is earned, not assumed. We combine multi-source data, preplanned anchors, rigorous multi-group modeling, and conservative decisions to make HCUs that travel. When they cannot, we say so and keep claims local.

These rules ensure we don’t water down scientific truth for the sake of a bigger N. It’s better to say “We found X in country A and Y in country B, but we can’t directly compare magnitudes because the scales differ” than to lump data together and draw potentially false conclusions.

Finally, we consider what could falsify the “same scale” ambition entirely. If we repeatedly find, for example, that scalar invariance fails across major language families *even after careful anchoring*, or that minor device tweaks create irreconcilable jumps in HCI, or that certain core items show DIF that we can’t adjust away and which *change our conclusions*, then our hope of a universal scale might be wrong. In that scenario, we would have to confine HCI to local uses or fundamentally redesign it. The take-home point: “*Same scale*” is earned, not assumed. We combine multi-source data, physical anchors, rigorous multi-group modeling, and conservative decision rules to make HCUs that travel. And when they cannot, we are honest about it and keep claims local.

8.1.5 Where we go next:

With the model-level requirements for “same scale” in view, we move from equations to lived diversity. 8.2 examines how language, display rules, development, aging, and social structure shape the signals we record—and how to model those influences so that differences in meters are not mistaken for differences in experience.

8.2 Culture and Age Effects

Fair measurement must work in bodies and communities that are not interchangeable. Language, cultural norms, climate, developmental stage, and aging each modulate how feelings are expressed or measured – potentially without changing the feelings themselves. This section maps the main sources of such variation and shows how HCI/HCU can accommodate them without losing comparability. When complete accommodation isn't possible, we'll see how to limit claims appropriately (e.g. only within-group conclusions).

8.2.1 Language and reporting styles

Problem: Words for distress and comfort don't translate perfectly. For example, words like "heavy-hearted" or "calm" carry culture-specific nuance. Some cultures encourage modesty in self-report and avoid extreme ratings, whereas others may use the full scale liberally. Public vs. private settings can change responses (people might downplay suffering if others will see their answers). All this means that a raw self-report number can't be naively taken as the same feeling across languages.

Response:

- Item crafting: We prefer concrete, low-idiom prompts that minimize translation issues. For instance, instead of asking "I feel sad," we might ask "Right now my body feels: [tense vs. relaxed]" or use simple phrases that have direct equivalents. We also use bipolar slider scales with short vignette anchors (e.g. one end described as "I feel very tense, like before an exam" and the other "I feel very relaxed, like after a hot bath") to give context in any language.
- Response-style modeling: We explicitly model tendencies like acquiescence (yea-saying) or extreme responding. In a CFA, we can include a latent factor capturing the person's general tendency to use higher ratings, or use IRT with group-specific item slope adjustments. By adding these as nuisance factors, we adjust scores for groups that, say, tend to use the middle of the scale more often. We also use MIMIC models to detect and correct reference-group effects (where a cultural group's idea of "5/10" might systematically differ).
- Bilingual bridges: When rolling out to a new language, we include a small sample of bilingual participants who answer in both languages. Their data help link the two language versions of HCI by seeing how the same person might give different numeric answers in each language. This helps us adjust thresholds so that a "7" in one language corresponds to the same latent F as a "7" in another.

- Partial invariance if needed: If despite all efforts, some self-report items still show scalar non-invariance (different intercepts/thresholds), we allow those to vary (free those parameters) and rely more on *non-verbal channels* (physiology, behavior) for cross-group comparison. For example, if the word for “anxious” just doesn’t work equivalently, we won’t base cross-cultural comparisons on it – we’ll lean on heart rate or cortisol which have more universal meaning.

8.2.2 Social norms and emotional display rules

Problem: Cultures differ in display rules – norms about when and how it’s acceptable to show certain emotions. In some collectivist cultures, overt anger or sadness might be suppressed in public, whereas in some individualist contexts, people are encouraged to express feelings openly. Thus, observational or behavioral indicators (like facial expressions, tone of voice) might vary not because of different feelings but because of different norms. Likewise, in self-reports, some might underreport negative feelings to save face.

Response:

- Context tagging: We annotate each experience sample (EMA) with context tags, such as whether the person was alone or with others, and whether the report was private or might be seen by someone (for instance, on a shared device). These tags are entered as nuisance variables Z_t . This allows us to adjust for context – e.g. if public reports are consistently toned down, the model can account for that when estimating F .
- Behavioral proxies: In cultures where self-report or overt expression is restrained, we put more weight on *indirect behavioral measures* of emotion. For instance, in tasks, we look at “stickiness” vs. “abort” rates (does someone persist or give up in a challenging task, which can indicate frustration or motivation) and social network dynamics (like whether someone seeks contact or withdraws after an exclusion, measured via phone data). These behaviors can serve as downstream validators of feeling that are less subject to self-report bias.
- Social-cost modeling: In the Queue System (LoF’s decision model), we incorporate social penalty parameters that are tuned per culture. For example, in a culture where expressing anger has high social cost, our models of decision-making (QS) will include a higher penalty for actions that cause visible anger. This ensures when we simulate or interpret behavior, we’re not pathologizing a culturally normative suppression as “lack of emotion,” but rather as a rational strategy under higher social cost.

- Avoid pathologizing differences: Low expressivity is not a “deficit” – it can be an adaptive strategy. We frame our analysis such that if a group is more emotionally reserved, we consider it a style, not evidence that they feel less. This perspective guides interpretation and model fitting (e.g. we wouldn’t assume a flat facial affect in a stoic culture means low F; we check other channels).

8.2.3 Climate, physiology, and device ecology

Problem: Environmental factors can systematically affect physiological signals. Heat and humidity increase skin sweat (conductance) regardless of emotion; altitude and fitness affect heart rate variability (HRV) baselines; people with darker skin or different skin properties might yield different photoplethysmography (PPG) readings for the same physiological event due to device optics differences. Also, devices sold in different regions might have slight firmware variations. In short, *where and how* you collect data can imprint on the numbers.

Response:

- Environmental covariates: We continuously record ambient temperature, humidity, altitude, light levels, etc., and include these as part of Z_t (nuisance inputs) for channels that are affected. For example, if Skin Conductance Level (SCL) is known to drift upward in hot weather, our model can subtract out the portion of SCL variance explained by temperature. We ensure seasonal or regional climate differences don’t masquerade as “happier” or “more anxious” signals.
- Device calibration and metadata: When we introduce a new device or firmware, we either do a common-person test (same participant wearing both new and old devices to find a mapping) or require a post-update re-anchoring session (repeat the anchor tasks after a firmware update to recalibrate). Also, every dataset point is tagged with device model/version, and we publish device version in metadata so others can be aware. If two sites use different devices, we won’t merge their data until we apply a linking function between the devices.
- Channel re-weighting: If we discover that a channel is unreliable or biased in a certain environment, we adjust its influence. For instance, in a tropical site, if galvanic skin response (SCL) is too noisy (everyone’s sweating all the time), we might down-weight SCL in the HCl factor model for that site. Meanwhile, we lean more on channels that remain stable, like HRV or pupil dilation or behavioral measures. Hierarchical Bayesian models allow us to “shrink” a problematic channel’s loading toward zero for that group, essentially letting the data say “this channel isn’t contributing here.”

- Drift alarms: We have automated monitors that look at the site's anchor results over time. For example, if every quarter we run a pain/relief anchor and suddenly one quarter the average results are off by >0.3 HCU from the historical average, that triggers an alarm. It could indicate device drift, sample change, or protocol issues. We then pause pooling data from that site until resolved (see 7.5.11 and 7.5.12 on drift handling).

8.2.4 Development: childhood and adolescence

Problem: Children aren't just "small adults" emotionally. They have evolving interoception (awareness of internal feelings), limited emotion vocabularies, and different patterns (e.g. their sleep cycles, which affect mood, are very different). Adolescents, on the other hand, have intense social and reward dynamics (peer acceptance, risk-taking) that color their affective life. If we use an adult-centric HCI, we might miss or misinterpret things in younger people.

Response:

- Age-tuned items: We swap out or rephrase certain self-report items for younger participants. For example, instead of asking a 10-year-old "I feel agitated," we ask "My stomach feels upset" or "I feel like I want to cry" – concrete sensations or actions that map to emotion. We also use simpler scales (smiley faces, etc., if needed for very young ages).
- Parent/teacher reports: For children, we add an observer-report component. A parent or teacher might regularly report on the child's apparent mood or behaviors (this is essentially another channel in the HCI model). We model the parent-report as correlated with the child's self-report but not identical – it gets its own error term and possibly a bias. Including this helps when kids can't articulate their feelings well.
- Sleep and dreams emphasis: Kids and teens have different sleep patterns (more REM in certain stages, adolescents often REM-deprived due to schedules, etc.). We put heavier weight on sleep-based indicators (like nightmares or deep sleep proportions) for younger people's affect balance. For example, if a child can't say they are anxious, their disturbed sleep might tell us. We also adjust anchors: for instance, a "mild pain" anchor might have a smaller effect in children because their range is different, or we emphasize a *social exclusion* game as a strong stimulus since peer rejection is very salient for teens.
- Developmentally valid tasks: We choose tasks that fit the age. For adolescents, we might include a peer evaluation game or an exclusion game ("Cyberball" or similar) to elicit social feelings, since teen affect is heavily social. We pre-register

expectations like “teens will have lower absolute horizon (they think less about far future) in social contexts.” That means when testing QS horizon effects, we account for developmental stage. We also explicitly test invariance across age bands (e.g. late childhood 8–12, adolescence 13–17, young adult 18–25) to see if the measurement model holds or needs age-specific parameters. If scalar invariance fails widely across age, we would avoid comparing, say, a 15-year-old’s HCU to a 30-year-old’s directly – instead, keep comparisons within age bands.

8.2.5 Aging and neurodegeneration

Problem: In older adulthood, baseline physiology shifts (e.g. blood pressure, sleep architecture), and many people take medications that affect mood or arousal (beta blockers dull heart rate responses, etc.). Neurodegenerative diseases (like Alzheimer’s or Parkinson’s) can blunt interoception (awareness of pain or emotion) and change how people report feelings (due to memory issues or apathy). If we ignore these, we might attribute differences to “affect” when they’re really physiological or cognitive constraints.

Response:

- Medication registry and adjustments: We keep a detailed log of medications for older participants. In analysis, we model drug-class effects on channels: e.g. if someone is on a beta-blocker, we expect lower HRV and dampened heart rate reactivity, so we adjust or include an interaction term for “on beta-blocker × HRV” in the model. We can produce both an “as-treated” HCU (their actual experience, including medication effects) and a “med-adjusted” HCU that attempts to factor out the medication’s influence for comparability.
- Alternate and passive channels: When self-report becomes less reliable (due to cognitive decline or even just different baseline), we lean more on behavioral and passive signals. For instance, we might use speech prosody analysis, facial micro-expressions (if consented to), or caregiver EMA reports as additional channels in HCI. If an older adult can’t remember how they felt in the morning, perhaps their step count, time spent active, or a prompt to a caregiver can fill in. These are modeled with appropriate error and not assumed to be perfect, but they add information.
- Longitudinal checks: We periodically re-check invariance within the same individuals as they age (e.g. annually). This can differentiate *true* affective drift (say their happiness truly declines) from *measurement drift* (say they start giving more neutral answers because of fatigue or a health issue). If we see that for an individual the relationship between channels is changing drastically over time, we

treat that carefully (it might indicate onset of a condition that changes their affect expression).

- Gentler anchors for frail populations: We don't subject a frail 80-year-old to a strenuous cold pressor if it's unsafe. For older or physically vulnerable groups, we use gentler anchor stimuli – e.g. a milder cold or a controlled breathlessness task instead of ice water, and perhaps a *shorter* social exclusion game with quick debrief. The idea is to still calibrate the scale, but within safe limits. If that means our ± 1 HCU definition shifts a bit (maybe for frail groups, ± 1 HCU is anchored on a milder stimulus), we document that in HCU versioning.

8.2.6 Social structure: poverty, safety, and time use

Problem: Socioeconomic and safety factors profoundly affect emotional life and how it's expressed. Someone living in poverty or under threat might have compressed choice sets (they simply have fewer options day-to-day, which can flatten affect range or decision patterns). They may also be continuously managing stress, which changes baseline autonomies (e.g. higher baseline cortisol). Their time use is different – long work hours, multiple jobs, caregiving duties – which can lead to *less* emotional variance (life becomes routine or survival-focused). If we compare such a person's HCI data to someone in a low-stress environment, differences might reflect context, not any fundamental difference in affect mechanism.

Response:

- Horizon and context in models: We explicitly model objective constraints like time scarcity. For example, we include in QS models a person's available free time, financial stress indices, etc., and treat their effective planning horizon (H) or feasible option set as *exogenously smaller* in high-stress contexts. This way, if someone shows "narrow" choices or low variance in HCI, we might attribute it to having no time for leisure rather than a lack of emotional capacity.
- Noise-robust data collection: In chaotic contexts, asking someone to fill a 10-minute survey might be unrealistic. We adapt by using wearables and brief EMAs that fit into busy lives. For example, a wearable can capture heart rate changes during a hectic day, and we ping a 1-question mood check when possible rather than lengthy forms. We also budget resources to ensure inclusion (e.g. providing participants with devices or data plans if needed) so that the sample isn't biased by who can afford tech.
- Interpretation guardrails: We train analysts and algorithms to avoid moralistic interpretations of flattened affect in hardship contexts. If we see lower variance in

HCI or subdued responses, we consider contextual constraints – not jumping to “this group is less emotional or more stoic by nature” but rather acknowledging the possible impact of external pressures. For reporting, we might note “Group X showed a narrower range of HCU, likely reflecting external constraints on their activities and expressions, rather than a reduced capacity for emotion.”

8.2.7 Culture-specific channels: music, ritual, and social pain

Opportunity: Not all cross-cultural considerations are problems; some are opportunities to use culturally salient stimuli as probes for universal responses. Across virtually all cultures, certain experiences evoke strong affect: music-induced chills, synchronous movement in ritual or dance, and experiences of social inclusion or exclusion. The forms differ (one culture’s folk dance vs. another’s religious ritual), but the underlying emotional push is shared (elevation, joy, connection, or pain from exclusion). We can exploit these as universal anchors or tasks that fit locally but measure globally.

Response:

- Universal anchors (sneak peek to 8.3): We incorporate tasks like physical pain, aesthetic chills, and social exclusion as part of our calibration battery precisely because they evoke biologically rooted responses. Pain is pain (with caveats). Indeed, Wager (2013) identified an fMRI-based “neurologic signature” of nociceptive heat pain—a distributed pattern that distinguishes painful from non-painful stimulation and tracks intensity—indicating a consistent neural footprint for physical pain. Music-induced chills recruit dopamine and arousal systems in convergent ways, and social exclusion reliably elicits distress across cultures. By using these as common denominators, we tie HCI to experiences with cross-population biological grounding. (More in 8.3 on how exactly we do this.)
- Local flavor, shared metrics: We allow cultural customization of stimuli while measuring the same outputs. For example, for an aesthetic “chills” anchor, each site can choose music that locally is known to give goosebumps (a church choir song in one place, a pop anthem in another). The *input* differs, but we measure the same *outcome*: piloerection (goosebumps via sensor or EMG), HRV spike, pupil dilation, self-reported chill intensity. As long as we capture the physiological signature, we’ve achieved a common anchor. Similarly, a social exclusion exercise might be framed differently (a ball-toss video game vs. a real-life scenario) but as long as it reliably triggers that hurt feeling (with corresponding ACC activation or HRV drop), it serves as a calibration point.
- Ethical use of social pain: We keep such tasks short and well-debriefed. A virtual ball-toss game (Cyberball) where someone is excluded for a minute or two can

induce a pang of social pain, but we immediately debrief (explain it was part of the study) and provide a positive resolution (an apology message or inclusion afterwards). We also always let participants opt out of these if they're uncomfortable. The goal is calibration, not distress.

8.2.8 Dreams across cultures and ages

Problem: Dreams are part of our affect regulatory system (as hypothesized in Chapter 10), but dream recall and content can vary widely. Culture influences how much people value or report dreams – some may consider them vital messages, others may ignore them. Age also matters: teenagers recall vivid dreams more often (and often emotional ones), whereas older adults might have more fragmented sleep and less recall. Work patterns and pharmacology (like antidepressants suppressing REM sleep) also change dream content and recall.

Response:

- Minimal dream channel: We include a “dream affect” channel but make it lightweight in contexts where recall is low or talking about dreams is taboo. Instead of asking for detailed dreams (which some might not provide), we might simply ask each morning: *Did you have a dream? If yes, how did it make you feel?* – offering a quick rating and maybe a tag like “was it about a loss, a success, being chased, etc.”. This yields at least a one-line affect rating of dreams without burdening those who don’t remember much.
- Polysomnography (PSG) subsamples: To validate that our interpretation of dream affect is consistent across groups, we sometimes do small lab studies with sleep recordings (PSG) in different cultures. For instance, confirm that when people say they had an intense negative dream, we see the known REM heart rate spike and that their next-day mood correlates inversely (the “dream as overnight therapy” hypothesis). If a culture under-reports dreams, we propagate that uncertainty by widening that channel’s error – we don’t assume no report means no effect.
- Adjust by age: For adolescents, dreams (especially nightmares or aspirational dreams) might be a larger part of emotional processing, so we weigh that channel a bit more for them. For older folks with poor REM sleep, we weigh it less. The model can include an interaction of age with the dream channel’s loading.

8.2.9 Equity and governance.

Principle: In seeking a universal measure, we must ensure the process doesn’t become a new source of unfairness or harm. Measurement itself can be intrusive or biased if done without community input. The guiding principle is: *the act of measuring should not*

adversely affect those measured. This is both an ethical stance and a practical one (distrust or harm will undermine data quality).

Practices:

- Community review and co-design: Before deploying HCI in a new cultural context, we engage local stakeholders. We have local advisors review the wording of items, the consent process, and the cultural appropriateness of tasks. In other words, translate first, not last – involve the community at the design stage, not as an afterthought. This often catches potentially insensitive or confusing elements.
- Privacy by design (on-device processing): Recognizing different norms around privacy, we ensure that any raw data like text or audio stays on the participant's device whenever possible. We only upload summarized or derived features (like “daily HCI score” or “speech sentiment score”). For example, if participants journal verbally, the recording is analyzed for sentiment *locally* and only sentiment scores are sent, not the recording. This protects participants and respects jurisdictions with strict data laws.
- Tiered consent: We allow participants to opt out of certain channels or tasks without leaving the study entirely. Someone might say, “I’m not comfortable with the dream journal part” – that’s fine, they can skip it and still contribute everything else. We design the study so that such opt-outs don’t invalidate all their data. This ensures people aren’t forced into sharing things they consider too private (like GPS data or intimate social info) just to be included.
- Transparency in diagnostics: We publish summaries of how the instrument performed in each group – e.g. which items were freed in partial invariance for a given culture, or what the channel weights ended up being for each demographic. By being open about “HCI worked less well on channel X for group Y, so we adjusted,” we invite critique and replication. If anyone suspects bias, they have the info to trace it. This is scientific honesty and also assures communities that if the tool isn’t working for them, we’ll see it and address it.

8.2.10 Decision rules for cultural/age transport

Given all the above adjustments, we still codify when we consider HCI/HCU truly comparable across cultures or age groups, versus when we report results separately. The rules closely mirror those in 8.1 but applied to these grouping factors:

- We pool data across cultures/age groups only if:
 - Configural and metric invariance hold (same structure and equal loadings across groups).

- Scalar/threshold invariance holds, either fully or via partial invariance/alignment with $\leq 25\%$ non-invariant items and no significant bias in means.
 - Anchor medians at each site/group are within 0.3 HCU of the overall reference. (If one culture consistently gives a larger or smaller HCU response to the standard pain anchor, that's a red flag – either adjust or don't pool.)
 - Device/firmware differences are either absent or accounted for (linked/standardized).
- Otherwise, we do not mix the groups for “law-level” claims. We keep comparisons within-group only and say, for example, “Invariance was not sufficient to compare absolute HCUs between teenagers and adults, so we only discuss trends within each group.” We also document in such cases which channels carried most weight in each group’s HCI (maybe teens’ HCI relies more on self-report, adults on physiology, etc.) and show that analyses with and without the problematic items yield similar conclusions within groups.

In essence, when in doubt, we err on the side of caution: treat groups separately and be explicit that LoF is being tested within a narrower scope. Only when we’re confident the scale is doing the same thing for everyone do we make broad statements.

8.2.11 Worked illustrations

Let’s ground this with a few hypothetical examples of how we apply these principles:

- Riyadh vs. Reykjavík: Imagine deploying HCI in Riyadh (hot climate, more conservative display norms) and Reykjavík (cold climate, more individualistic expression). In Riyadh, we find skin conductance (SCL) is very noisy due to heat and people tend to under-report distress in public settings. In Reykjavík, SCL is stable but people use the full range on surveys. Our solution: we down-weight SCL in the Saudi data (and instead rely on HRV, pupil dilation, etc., which are less climate-sensitive), we encourage more private EMAs in Riyadh (having people report anonymously or when alone to get candor), and we use the same anchor tasks (a cold pressor and a relief stimulus) in both sites to calibrate. We find that after these adjustments, metric invariance holds and only a couple of survey items needed threshold adjustments for Riyadh. We then feel confident comparing (with partial scalar invariance) certain results, while noting that absolute values might still carry ± 0.2 HCU uncertainty between the sites.

- Middle-school cohort: We run HCI in a middle school. We use language-light, body-focused items (“felt left out” instead of complex emotion terms) and emphasize peer interaction tasks – e.g. a game where feedback from peers can raise or lower your mood – to engage their social context. We include a strong social exclusion anchor (perhaps a brief “no one passes you the ball” online game) to gauge their negative HCU, and we keep dream questions minimal (maybe just ask for a bad dream count). We pool data within the 11–14 age range but do not compare their HCI means to adults, since scalar invariance fails when including adults (adolescents might have inherently higher volatility, etc.). We report findings relative to their own baseline or vs. an intervention, but not “teens vs. adults” on the HCU scale unless invariance passes in a separate analysis.
- Memory clinic (older adults with cognitive impairment): We run HCI in a memory clinic for early Alzheimer’s patients. Recognizing issues, we incorporate caregiver EMA (someone logging the patient’s apparent mood), sleep tracking, and simple behavior tasks (like reaction to a favorite song) rather than expecting detailed self-reports. We use gentle anchors only – perhaps a mild hand cold test instead of full cold pressor, or a slow breathing exercise to anchor calm. We don’t do a full-on Cyberball exclusion because it could distress them disproportionately. We adjust each person’s HCU for their medication load (since many are on drugs affecting the nervous system). We heavily rely on longitudinal invariance here: each patient is their own reference over time, and we see if their HCI can still track within-person changes reliably. We might not even attempt to pool across individuals here, given heterogeneity; instead, we ensure each individual’s ledger is calibrated and then later see if any group patterns emerge.

Takeaway: Culture and age do not invalidate a shared affect scale; they stress-test it. By designing HCI for invariance (linguistically and culturally neutral where possible), anchoring it in universal human experiences (pain, relief, joy), redistributing channel weights as needed, and being transparent about diagnostics and uncertainties, we aim to keep HCI and HCU honest wherever they travel – and honest about where they *don’t* travel well. The Law of Fairness as a concept is universal, but our measurements must earn their universality through this careful process.

8.2.12 Where we go next:

Having mapped culture and age effects, we now anchor the scale in shared human responses. 8.3 introduces a practical triad—nociceptive pain, aesthetic chills, and brief social exclusion—that gives us portable, ethically bounded yardsticks so a unit of HCI means the same thing when carried from one site to another.

8.3 Universal Anchors (Pain, Chills, Social Exclusion)

To keep HCU as a shared currency across people, we explicitly calibrate it using experiences that are as universal as possible – stimuli that almost anyone can feel, regardless of language or background. Three families of stimuli meet our criteria for universality, controllability, and ethical use: nociceptive pain, aesthetic chills, and social exclusion. Each of these elicits a reliable, time-locked change in affect with signatures across multiple channels (self-report, physiology, brain activity, behavior, and even sometimes dream content). In this section, we detail the protocols for each anchor, what HCI/HCU patterns we expect them to produce, safety limits (we must do these ethically), and how we use these anchors in both lab and field studies as yardsticks.

8.3.1 Design criteria for universal anchors

To choose an anchor stimulus, we consider several design criteria:

- Universality: The stimulus should reliably evoke an affective response in most healthy people across different cultures. It should not depend on specific cultural knowledge. For example, a piece of instrumental music can evoke chills in many populations, whereas a stand-up comedy clip might not be funny cross-culturally.
- Parametric control: We need to be able to titrate the intensity and duration of the stimulus to produce graded changes in HCU. For instance, we can make cold water colder or keep a hand submerged longer to increase pain; we can pick a song segment known for a mild chill versus a full-blown frisson.
- Multichannel footprint: The anchor should cause changes in at least two different types of measures – say, both a self-report and a physiological change. Ideally, it touches many channels: e.g. pain makes your heart rate and skin conductance spike and also makes you report feeling bad – that convergence is gold for us.
- Ethics: The stimulus must be safe, with only transient discomfort, and be something an ethics board approves. Participants must be able to stop at any time, and debriefing should remove any lasting negative effect. This rules out anything truly traumatic or harmful.
- Transportability: It should be feasible to administer in different settings, including possibly at home (micro-anchors). We prefer anchors that don't require extremely specialized equipment so that a clinic in one country and a lab in another can both do it. For example, a bucket of ice water is low-tech and can be done anywhere; an fMRI task is high-tech and site-specific (we use the former as an anchor, the latter as outcome measurement instead).

With those in mind, we identified three anchors that complement each other on the HCU scale: one for negative/somatic affect (pain), one for positive/aesthetic affect (chills), and one for negative/social affect (exclusion). Each covers a different “domain” of human experience, helping ensure our scale isn’t narrowly tied to just one kind of feeling.

8.3.2 Anchor A — Nociceptive Pain (Cold Pressor)

A participant immerses their hand in near-freezing water during a cold pressor test. This controlled pain stimulus is used as a universal pain anchor to calibrate the negative end of the HCU scale across individuals. By defining a standard integrated impact over a fixed window (e.g. -1.0 HCU for the cold pressor window), we ensure that “-1 HCU” represents a comparable real-world burden for everyone.

Rationale: Pain is a nearly universal experience and arguably one of the most fundamental negatives across species. It triggers strong autonomic responses (like heart rate and blood pressure spikes) and robust brain activation in areas like the insula and anterior cingulate cortex (ACC). Importantly, it doesn’t require translation: sticking your hand in ice water hurts whether you speak English or Hindi. The dose-response is reliable (colder or longer = worse pain). Thus, pain provides a common reference for the low end of the affect scale.

Core protocol (lab setting): We often use the cold pressor test:

- Stimulus: The participant immerses their forearm or hand in ice-cold water (typically 0–2 °C). We set an upper time limit (e.g. maximum 90 seconds) but instruct participants to remove their hand whenever it becomes intolerable.
- Measures: We continuously record physiological data: heart rate variability (HRV, especially RMSSD or high-frequency components), skin conductance level (SCL), pupil dilation. We also ask the participant to give a very short rating (0–10 scale, or squeeze a pressure sensor) every 15–30 seconds to gauge pain level or affect. In some sessions, we include neuroimaging (fMRI or EEG) on a subset of participants to see the neural response, and we note any behavioral signs (like facial grimace, or whether they withdraw early).
- Phases (Windows): The procedure has a Baseline (~2 minutes of rest, hand in room-temp water or no water), then Immersion Ramp (first ~30s of the hand going into ice water, pain rapidly rising), then a Plateau (e.g. last 30s of whatever duration they manage, where pain is near max tolerable), and then Recovery (~2 minutes after removing hand, where the sensation normalizes and often there is relief).

Expected HCl/HCU signature: We expect a sharp drop in HCl as pain is introduced. During the ramp and plateau, the composite HCl (which aggregates negative feelings) should drop significantly into the negative. In recovery, HCl should rebound – possibly even above baseline for a bit, due to a relief/counterweight effect (the feeling of relief is a positive swing).

Concretely, we operationally define -1.0 HCU in our global scale as the median drop in latent affect *at the end of the plateau* for a reference group of participants. For example, if across a reference group the integrated HCl over a prespecified cold pressor window yields a median cumulative value of -1.0 HCU, we use that as the anchor for the negative unit. This ties the abstract unit to a real sensation: “one HCU down = what a typical person feels at the worst moment of a short ice-bath of the hand.”

We also verify dose-response: longer immersion or slightly colder temperature yields larger HCU drops (monotonic). If someone only lasts 30 seconds vs. another lasts 90, the one who lasted longer likely had a bigger HCU drop (or higher pain tolerance, which we interpret carefully). This monotonic relation is used to check consistency across sites – if one lab’s “90s in ice” yields -0.8 HCU and another’s yields -1.2 HCU, we investigate why (subjective pain reports, selection differences, etc.).

Safety/ethics: We have strict guidelines:

- Exclusion criteria: Participants with conditions like Raynaud’s disease (extreme sensitivity to cold), peripheral neuropathy, or cardiovascular problems are not asked to do cold pressor.
- Abort ability: We emphasize they can pull their hand out at any moment with no penalty – their data is still useful. We observe closely and remove the stimulus if someone shows any concerning signs (like a vasovagal near-faint).
- After the cold pressor, we provide a warm towel or compress to alleviate discomfort quickly, and we monitor for any delayed vaso-vagal responses (faintness, etc.).
- For more frail groups (older adults, etc.), we often use a milder version (water at, say, 5°C , or a shorter exposure) to avoid undue stress. We might also impose a shorter max duration (30s) for them.

Micro-anchor (field version): We can’t dunk people’s hands in ice at home easily, but we have simpler approximations for in-the-wild calibration:

- For example, a commercial “cold pack” (like those chemical cold packs) applied to the forearm for 20 seconds can induce a mild pain.

- Or a thermal coin device that briefly cools a small area of skin. We pair this with wearable measurements (e.g. a smartwatch capturing HRV and perhaps a phone prompt “rate your discomfort 0–10”). This typically yields a modest negative HCU deflection, say -0.3 to -0.5 HCU. We wouldn’t define the whole scale on this, but it’s useful for recalibration between lab visits. For instance, if someone’s device is replaced, doing a quick cold-pack test at home and seeing they get -0.4 HCU now vs. -0.5 earlier gives a check that things are consistent.

In summary, the pain anchor establishes the negative calibration point: it gives everyone an experience of “this is pretty bad” in a controlled way, and pins the HCU scale’s lower bound (within ethical limits). Next, we want an anchor for a *positive* burst of affect.

8.3.3 Anchor B — aesthetic chills (music or art)

Rationale: Not all intense experiences are negative. The frisson or “chills” one gets from a powerful piece of music, a moving speech, or a breathtaking scene is a spike of *positive* affect that appears nearly universal. People across cultures report goosebumps or shivers down the spine during certain aesthetic experiences (particularly music). Physiologically, chills correlate with distinct patterns: piloerection (hair standing on end, measurable as goosebumps), a spike in HRV and sometimes a lump-in-throat or teary-eyed response. Brain-wise, chills activate reward circuits (like the ventral striatum and vmPFC) similarly to other pleasures. Importantly, it’s a safe high – unlike taking a drug or something – and has no adverse aftereffects. Thus, aesthetic chills serve as a good positive anchor. Not everyone gets them strongly, but a large subset do, and we can work with that.

Core protocol (lab):

- Stimulus: We use music (most common trigger for chills). We typically ask participants beforehand for a song or piece that *consistently gives them chills*. Additionally, we have a library of known frisson-inducing pieces (like certain climaxes in classical music, or emotionally intense film scores). In the session, we play a segment of music that includes a “cold” control section (no big climax) and then the chill section where the big emotional swell is. Sometimes we also use spoken word or video if culturally relevant (e.g. a moving poem or scene).
- Measures: We attach a piloerection sensor – yes, there are such things: either a Goosecam (camera and image analysis of skin) or an EMG that detects the tiny muscle movements when hairs stand up. We record HRV, pupil size, and we ask participants to press a button when they feel a chill or goosebumps. We might

also use EEG or fMRI in subsamples to see the neural response, but the key measures are peripheral and self-report.

- Windows: We define sections: a “cold” spot (no chills expected, acts as baseline control within the piece), then a build-up, then the peak (where chills usually happen), and an afterglow period. We compare physiology and HCl during peak vs. cold segments.

Expected HCl/HCU signature: During the chills, we expect a transient positive spike in HCl – e.g. a jump of +0.5 HCU or more, but for a short duration (a few seconds). On composite, this might appear as a little “blip” of positive affect. Many report a lasting positive mood after a strong chill, so we also watch for an afterglow elevation – maybe their baseline HCl is a bit higher for a minute after. These anchors complement the pain anchor; one gives a sustained negative, the other a brief positive surge.

We often use chills as a calibration to check the upper range of HCU, in tandem with an analgesic relief anchor (because giving someone pain relief is another positive, see below). Chills are like a natural “reward” test. We also use it as a neutralization check after negative blocks: for example, after a pain block, playing their favorite chill-inducing music helps see if their HCl rebounds to baseline or overshoots (if it overshoots, maybe our counterweight model needs adjusting).

Safety/ethics: Pretty easy here – listening to music is low risk. We do screen for any issues like epilepsy (if using flashing videos) or extreme emotional sensitivity. Also, respect content: if someone’s chills piece is sacred or tied to personal memories, we ensure the setting is respectful (e.g. if it’s a national anthem or prayer, we treat it appropriately).

- We keep volume at safe levels to avoid any hearing damage.
- We ensure the content is not offensive in that culture (we let them pick or vet it).
- And obviously, we wouldn’t use this if it somehow triggered negative emotions (usually chills are positive, but say someone gets chills from a horror movie – we might skip that because it’s more fear than awe we want).

Micro-anchor (field): For everyday calibration, we can use a short playlist of the participant’s chosen “chill” segments (maybe 2–3 minutes total). They put on headphones or we instruct them to sit and listen. Their wearable logs HRV and pupil, and they have a button to press on a phone app when they feel a chill. This can validate that their device captures a +HCU response outside the lab. It’s especially useful in populations where giving a drug or doing a big lab setup for a positive stimulus (like giving actual medication for euphoria) is impractical – music is accessible to all.

We primarily use aesthetic chills to validate the upper range of the scale: ensure that what we call “+1 HCU” actually correlates with a genuinely positive subjective experience across people, including those who may not be able to get pharmacological interventions (like certain patients).

8.3.4 Anchor C — social exclusion (ostracism)

Rationale: Humans are deeply social, and social pain (like being excluded or rejected) can be as powerful as physical pain. Neuroscience shows overlapping neural circuitry between social pain and physical pain (the dorsal ACC and anterior insula activate in both). Indeed, even a physical analgesic (acetaminophen) has been shown to reduce social pain responses, providing direct evidence of a shared neurochemical basis (DeWall, 2010). So, to cover the domain of *social* negativity, we use a standardized ostracism scenario. It's reliably distressing in a mild way, works for children up to adults, and can be ethically handled. This helps us see how HCI responds to social hurt, complementing the bodily hurt of the cold pressor.

Core protocol (lab): The classic paradigm is Cyberball – a computer game where the participant believes they're tossing a ball with two or three other players. After some inclusion, the others stop throwing the ball to the participant (excluding them). Another approach is a brief real-life scenario: e.g. two confederates in the lab start ignoring the participant in a conversation. We typically prefer the virtual game for consistency.

- Stimulus: We program a session of Cyberball with a few minutes of inclusion (ball tossing equally) and then a block of, say, 2–3 minutes where the participant is completely excluded (they never get the ball). In some studies, we include a reconnection phase at the end where we explain or bring them back in, to see recovery.
- Measures: We capture HRV, SCL, pupil dilation continuously; we ask occasional EMA like “Right now, I feel left out” on a scale of agree/disagree. We record behavioral data from the game (did their throwing behavior change, though in Cyberball they have little control). In fMRI versions, we'd see dACC and anterior insula. Also, we might include performance measures if combined with a task (like solving a puzzle alone vs. others).
- Windows: Key windows are Inclusion baseline, Exclusion onset (the first moments they realize “hey, I'm not being included” – often the worst feeling), Sustained exclusion (the remainder of that period), and then Recovery/Reconnection.

Expected HCl/HCU signature: At the moment of exclusion onset, HCl should drop (sudden negative spike, maybe -0.5 HCU or more) due to the emotional pain of rejection. Some participants may partially recover or stabilize if they cope or rationalize ("maybe it's a glitch"), but overall negative affect remains elevated during ostracism. When we reconnect or debrief (tell them "actually, those players were just computers, it's part of the study, you didn't do anything wrong"), we often see HCl go up – a relief or positive "re-inclusion" effect.

This provides a social domain anchor that complements the somatic pain and aesthetic pleasure anchors. It ensures that if a person's life ledger is negative, it's not just because they felt physical pain, but also we're capturing social pain equivalently.

Safety/ethics: Social exclusion can have real emotional impact, so we handle it with care

- We keep exclusion brief (a few minutes) and mild. We don't, say, have their real friends shun them, which would be unethical. It's a contrived scenario with strangers or computer players.
- We pre-screen participants for vulnerability. Someone with severe social anxiety or recent trauma of exclusion we might exclude from this task, or get additional consent.
- Immediate debrief and repair: Right after the task, we explain the purpose ("this was to simulate feelings of exclusion which many people find unpleasant; it wasn't real; everyone gets excluded as part of the design"). We also include a "reconnection" interaction – maybe an experimenter comes in friendly, or we have the computer players send a nice message in the end, etc., to ensure we don't leave someone in a negative state.
- We offer an escape option even mid-task: if they get too upset, they can quit. And we often include a warm-up inclusion game first so it's not a shock when exclusion happens (some protocols do an include-exclude-include sequence).

Micro-anchor (field): We can simulate a tiny exclusion in daily life, though this is trickier ethically. One approach (opt-in only) is having a familiar chat app conversation where for a short window, the participant's messages don't get replies (simulated delay). After a minute or two, the "lag" is explained and resolved. During that time, we capture their heart rate or ask an EMA like "I felt left out in that moment" on the phone. This is delicate, so we only do it with explicit consent and if participants are comfortable (and we make it very short). The goal is to see if their wearable picks up a similar pattern (e.g. heart rate increase, etc.) when socially "pinged."

Together, these three anchors – pain, chills, and exclusion – form a triad. Each hits a different facet of affect, and each is something humans universally experience in some form. By tying the HCU unit to events almost everyone can understand (“that feeling of hurting yourself on ice,” “that goosebump feeling from music,” “that sinking feeling of being left out”), we give our scale a fighting chance at meaning the same thing across different lives.

8.3.5 Putting it all together: the anchor triad

Using all three anchors in a cohort gives us broad coverage and cross-checks:

- Somatic negative (pain): a robust, non-linguistic negative experience with strong autonomic signature. Defines the lower bound of HCU.
- Aesthetic positive (chills): a natural pleasure response with a distinctive physiology (piloerection, HRV spike). Helps define the upper bound and tests if positive counterweights work.
- Social negative (exclusion): a core human aversive experience tied to social context. Ensures our scale captures social pain, which might otherwise be underrepresented.

Together, these allow us to triangulate HCU. If one person’s HCI drops –0.8 from pain and another’s drops –0.8 from pain, we assume similar pain feeling. If their exclusion drops differ wildly, we investigate (maybe one had coping strategies). It also lets us test the Queue System (QS) predictions: e.g. QS suggests after a big negative, people seek a counterweight. We can see after exclusion, if giving a reconnection (positive social input) boosts HCI more than a neutral message – that’s a QS prediction. Or after pain, does hearing a chill-inducing song bring HCI back up more than a neutral song? These anchors thereby not only calibrate the scale but also serve as mini-experiments for our theory.

8.3.6 Parametric titration and dose-response

We design each anchor to be tunable so we can measure dose-response curves and ensure comparability across sites:

- Pain: We can adjust the temperature or duration of the cold pressor. Colder/longer should produce more negative HCU (to a limit). We expect a roughly monotonic relationship: e.g. 0°C for 60s might be –1.0 HCU, whereas 2°C for 30s might be –0.3 HCU. Each site can run a few variations to confirm they see the expected gradient.
- Chills: Not everyone gets chills from the first piece of music. We often present multiple candidate clips, or have an adaptive algorithm (choose the one they

rated highest during screening). We also note that the amplitude of the HRV or piloerection correlates with the intensity of HCU spike – we want to ensure a good one to represent +1 HCU. If needed, we have them identify not just one but a couple of “most chills” moments and perhaps repeat on another day to verify consistency.

- Exclusion: We can vary the percentage of exclusion (partial vs. complete) and the duration. We try to monitor the *first derivative* of HCl – if it’s plummeting too fast, that’s likely beyond -1 HCU for them and we might cut it short. Also, we might test a “dose” of reconnection: e.g. a short apology vs. a very warm reconciliation, to see if that neutralizes the negative fully or overshoots into positive (which informs our model of counterweight).

By having parametric control, we can create calibration curves for each site: e.g. how much does 10 seconds more of ice equal in HCU change, or how big a chill response corresponds to 0.5 HCU. We compare these across sites. If one site consistently shows weaker responses to all anchors, perhaps their participants are just more stoic or our equipment was less sensitive there – that might indicate a scaling adjustment needed.

8.3.7 Cross-site implementation checklist

To standardize anchor use across all labs in our consortium, we maintain a checklist:

- Stimulus library: A shared repository of vetted stimuli: for music, we have several tracks that have induced chills in prior research plus allow sites to add local music (with meta-data on which physiological markers responded). For Cyberball, we share the exact script/program parameters to ensure consistency. Essentially, every site has access to the same “anchor kit.”
- Hardware pack: We provide a recommended set of sensors: e.g. an HRV-capable chest strap or wrist device, a simple eye-tracker or high-frequency camera for pupil size, a finger GSR sensor (when climate permits), and a goosebump detection method (like a dermal camera or even a simple hack like an arm-mounted camera). Not every site can do fMRI, but those who can will at least follow similar timing if they do.
- Scripts and timing: We distribute scripts for each anchor task – e.g. a timed sequence for Cyberball (how many throws before exclusion, how long, etc.), or a standardized cold pressor protocol with timestamps for when to log events. EMA prompts are prepared in local languages, synced with events (like “Rate your pain now” pops up at 30s into cold pressor).
- Anchor panels: Each site is asked to run a small anchor panel (10–20 local participants) initially and annually. They administer all anchors and share

- anonymized summary stats (like distribution of HCU changes for each anchor). This populates a database to monitor any drift or differences. For example, if one site's pain anchor median is significantly off from others, we flag that.
- Version control: We treat any change in anchor procedure like a version update. If we change the cold pressor to using ice packs instead of water, that becomes anchor version 2.0 and we publish a note on how to convert old data to new (if possible). We log device firmware and software versions as well. If someone updates the Cyberball game (maybe graphics changes), that's noted. The idea is to ensure reproducibility and that any differences can be traced to their source.

8.3.8 Analysis plan and decision rules for anchors

When analyzing anchor data, we focus on a few key outcomes and criteria to decide if an anchor “worked” at a site:

- Primary outcomes: The main metrics are the HCU change during the anchor's peak (e.g. how many HCU down at pain max, how many up at chill moment) and the recovery trajectory (how quickly it bounces back). We also often compute the area under the curve (AUC) for the HCl change – basically, the total impact over time – as a measure of sensitivity (how much “signal” we got from the anchor).
- Convergence test: We set a rule that an anchor is only valid if it triggers *converging evidence* from multiple channels. For instance, if someone reports pain 8/10 but their physiology hardly changes, we'd be suspicious. So we require at least 2 channels to corroborate the HCl shift. In pain's case, typically self-report and either heart rate or conductance must show the effect; for chills, maybe goosebumps and HRV both respond; for exclusion, maybe both self-report and heart rate or pupil dilate with distress.
- Transport pass (site validity): When a new site runs the anchors, we check if their anchor effects are within 0.3 HCU of the pooled reference median for those anchors. If a site's anchors are consistently weaker or stronger, it suggests something's off (either procedural or population difference). In such case, we might recalibrate (adjust that site's HCU scaling) or decide not to pool that site's data until resolved.
- Counterweight check: We also examine if after a negative anchor, the positive anchor restores HCl towards baseline (or after a positive, whether things settle without overshoot). For example, after the cold pressor (pain), if we then give a chills stimulus, does HCl *approximately* return to baseline without overshooting? If it *overshoots* (goes above baseline significantly), maybe our +1 calibration is too

potent or our model of how counterweights sum is off. We incorporate this into QS model validation.

8.3.9 Special populations for anchors

We adapt anchor use for various groups to ensure everyone can participate safely:

- Children: For kids, we shorten and soften everything. A child might only do a 10-second cold pack on the hand instead of a 60s ice bath. We might use a cartoon version of Cyberball (animated characters tossing a ball) to make it less distressing and clearly a game. For chills, we can use an upbeat animated movie song if that's what gives them frisson. The key is to evoke some affect without crossing ethical lines. (Children can be surprisingly engaged by the tasks if framed as games.)
- Older adults / frail individuals: We avoid intense pain for them – maybe a mild cold pressor (slightly cool water) or even just a guided slow-breathing as an anchor for calm rather than stressing them. Social exclusion is shortened; and we emphasize reconnection because, say, an older person with cognitive impairment might take exclusion very hard unless we reassure them quickly.
- High-stigma or vulnerable contexts: In some cultures or situations, even a simulated exclusion might be too upsetting or misunderstood. We adapt by making it very evidently artificial (e.g. “sometimes messages arrive late; you will experience that” to simulate a mini exclusion, so it doesn’t feel like genuine social rejection). We ensure any EMAs about it are private. If any anchor could inadvertently shame or embarrass (like visible goosebumps sensors), we handle with utmost privacy.

8.3.10 Potential failure modes and remedies for anchors

We've encountered or anticipated a few things that can go wrong with anchors:

- No chills responders: Some people just don't get chills from music. If a significant fraction of a sample has zero response, we have backups. We consider using “awe” stimuli (like epic landscape videos or a meaningful ritual) which might trigger a similar emotional high. If one site finds that goosebumps are rare (maybe due to climate or genetics), we rely on *subjective* chills reports plus HRV/pupil. The anchor still works, we just measure it differently.
- Exclusion backfires with anger: Occasionally, instead of feeling sad, an excluded participant gets angry at the others. That's still negative valence (HCl goes down) but the profile might differ (anger might increase skin conductance more, for

instance). We note such cases – it's not a failure per se, just a different flavor of negative. If someone reacts with irritation rather than hurt, we still count it as negative HCU, but we ensure the recovery might need a slightly different handling (anger might not be fully “neutralized” by a simple apology – maybe they need a bit more cooling off). We shorten the exclusion if we see signs of strong anger to avoid any ethical boundary.

- Pain ceiling issues: People's pain tolerance varies. Some might pull out in 10s, others last full 90s. If too many hit the *ceiling* (e.g. a lot of people go full 90s and say “I could do more”), maybe our water isn't cold enough; conversely, if many can't last 20s, maybe it's too cold or too harsh. We adjust on the fly during piloting. The rule is safety over symmetry: it's okay if one site's -1 HCU is defined at 60s while another's at 90s, as long as both delivered a clearly significant pain.
- Logistical issues: e.g. a site lacking a certain sensor. If a site can't measure goosebumps, we may proceed with chills using just self-report and heart rate but with caution. Or if someone is on a medication that blocks physiological response (like beta blockers dull heart rate), we focus on their self-report and maybe facial expression for anchors and mark that their physiological channels won't show much.

Takeaway: The trio of pain, chills, and social exclusion gives us a portable toolkit to calibrate our HCU in units grounded in real feelings. These anchors are carefully chosen to be *human* (not culture-specific), tunable, and ethical. By ensuring one person's “ -1 HCU” was an ice-water pain and another's was the same sort of pain, and similarly tying $+HCU$ to chills and closure of social pain, we create a shared yardstick. This way, when we talk about how much better or worse a day felt, or balance a life's ledger, we're doing so on a scale that *travels* – from lab to lab, culture to culture – and stands on tangible experiences rather than abstract ratings.

8.3.11 Where we go next:

The anchors let us speak the same hedonic language across contexts; the next step is to certify that language statistically. 8.4 lays out the formal tests—configural → metric → scalar invariance, DIF checks, linking designs, and preregistered thresholds—that decide when pooling is justified and when it is not.

8.4 Research Notes: Configural → Metric → Scalar Invariance

This section is a hands-on playbook for testing whether our HCI/HCU measurements truly live on the same scale across different groups (languages, sites, ages, devices). It's about the statistical methods and pre-registered decision rules we use. Researchers can use this as a checklist when analyzing multi-group data to determine: can we treat these groups as comparable (and pool their data or compare means), or do we need to adjust/stratify?

8.4.1 The three rungs of invariance (what each means and why it matters)

- Configural invariance: The *pattern* of factor loadings is the same across groups. In other words, each group has the same basic model – the same indicators are associated with the latent factor, though the actual loading values may differ.
What it buys us: We can say the construct (HCI) exists in each group and has the same conceptual meaning (e.g. all groups distinguish between positive and negative affect in the same way). Without configural invariance, we're comparing apples and oranges.
- *How to test/model:* Fit a multi-group model where each group's loading matrix Λ_g , intercepts v_g , and residuals Ψ_g are free, but we ensure the same indicators load on the factor in all groups (no group has an extra factor or a completely different pattern). Good fit here (CFI, RMSEA acceptable) is step 1.
- Metric invariance: The *loadings are equal* across groups ($\Lambda_g = \Lambda$). This means one unit change in the latent factor corresponds to the same change in an indicator in all groups. *Benefit:* With metric invariance, we can compare relationships involving the factor across groups – e.g. if HCI correlates with some outcome, those correlations are meaningful to compare because a unit of HCI meant the same “amount of change” in each group’s measurements. However, we still wouldn’t compare factor means because each group might have a different intercept.
- *Test:* Constrain all λ ’s equal and see if model fit drops significantly (using ΔCFI or chi-square difference). If the change in fit is within a small tolerance (e.g. $\Delta\text{CFI} < 0.01$), we say metric invariance holds.
- Scalar invariance: Both loadings and *intercepts/thresholds* are equal across groups ($v_g = v$, and for ordinal items, thresholds $b_{\{k\}} = b_{\{kc\}}$). This means that the zero of the latent factor has the same meaning and the observed scores align on the same scale. *Benefit:* Now we can compare latent means (and thus things like absolute HCU values or ledger totals) across groups. Scalar invariance is needed to say “Group A’s HCI=5 is same level of happiness as Group B’s HCI=5.”

- *Test:* Constrain intercepts equal and again check fit changes. If fit holds (or if any drop can be mitigated by allowing a few intercepts to differ = partial invariance), we proceed with comparing means.

(Note: We rarely require strict invariance – equal residual variances – because differences in residuals often just reflect minor differences in measure reliability per group, which doesn't necessarily bias factor means. For HCI, as long as metric and scalar hold, that's usually sufficient.)

If scalar fails massively, it's a show-stopper for comparing means. We then either have to do partial invariance or just not compare those groups' means at all.

8.4.2 Model forms (CFA vs. ordinal IRT for HCI)

We generally use a mix of continuous and ordinal indicators in HCI. For transparency, here are the forms:

CFA (continuous indicators): As earlier in 8.1, for each channel k in group g : $y_{\{g,t\}} \sim N(\mu_{\{g,t\}}, \sigma^2_{\{g,t\}}) = v_{\{k,g\}} + \lambda_{\{k,g\}} F_{\{g,t\}} + \Gamma_{\{k,g\}} Z_{\{g,t\}} + \varepsilon_{\{g,t\}} \sim N(0, \Psi_{\{k,g\}})$. This covers things like physiological measures (which we might treat as continuous after some transformation). In multi-group CFA, to test invariance we do:

- Metric step: set $\lambda_{\{k,g\}} = \lambda_k$ for all groups (all loadings equal).
- Scalar step: additionally set $v_{\{k,g\}} = v_k$ (all intercepts equal). We always include nuisance regressors $Z_{\{g,t\}}$ (like arousal covariates, etc.) in each group as needed.

IRT (ordinal indicators): For questionnaire items or ratings on scales, we use IRT (graded response model). For group g : $P(y_{\{g,t\}} \geq c | F_{\{g,t\}}) = \text{logit}^{-1}[a_{\{k,g\}}(F_{\{g,t\}} - b_{\{k,g\}})]$. Here $a_{\{k,g\}}$ is analogous to a factor loading (discrimination slope) and $b_{\{k,g\}}$ are the thresholds between response categories. Invariance steps:

- Metric: set $a_{\{k,g\}} = a_k$ (equal slopes).
- Scalar (threshold): set $b_{\{k,g\}} = b_{\{k,c\}}$ for all c (equal thresholds). If full threshold invariance is too strict, we do partial (some thresholds free) or alignment.

In practice, we often use software that can handle both together (e.g. treat some items as ordinal, some as continuous). The key is always include relevant nuisance covariates (like Z for climate, device, etc.) in the model per group to account for known differences, so invariance is tested on the *core affect factor*, not trivial differences.

8.4.3 Fit targets and thresholds (pre-registered)

We pre-specify what counts as “good enough” invariance in terms of fit indices:

- Global model fit: For overall configural model, we aim for $CFI \geq 0.95$, $RMSEA \leq 0.06$, $SRMR \leq 0.08$ in each group or pooled model. These are conventional cutoff guidelines for a good factor model fit. If we can't achieve those even configural, something's wrong with our measurement model to begin with.
- Invariance constraints impact: We accept differences like $\Delta CFI \leq 0.010$ and $\Delta RMSEA \leq 0.015$ when adding constraints (metric or scalar). That means if forcing equality of loadings drops CFI by less than 0.01, we consider metric invariance supported. We also look at chi-square difference but fit indices are more practical with large N.
- IRT-specific diagnostics: We look at item-level fit statistics like $|S - X^2|$ (an item fit index) and expect them to be non-significant ($p > .05$) after adjusting for multiple testing. We also compare item information functions across groups (without relying on any required visual) to check whether items behave similarly. If many items show misfit, that hints at non-invariance or multidimensionality. Also, we plot item information curves per group to see if they overlap; large discrepancies mean trouble.
- Bayesian approach: Sometimes we use Bayesian CFA/IRT. There, we want posterior predictive p-values ~ 0.5 (model replicates data well) and we check credible intervals for differences in loadings between groups – they should include 0 and be narrow if invariance holds. We might set priors that mildly favor invariance to stabilize estimates.

If an invariance step fails these targets, we don't just give up – we move to partial invariance or alignment as described next. All these criteria are pre-registered in our analysis plans, meaning we decide them in advance so we're not fiddling post-hoc to claim invariance.

8.4.4 Partial invariance and alignment (rescue strategies)

When full scalar invariance fails, we attempt partial invariance or alignment:

- Partial invariance: Free the minimum necessary intercepts or thresholds to improve fit. Practically, we look at modification indices: e.g., it might say "Item 5's intercept differs in Group 2." If that item alone has a large misfit, we allow item 5 to have its own intercept in Group 2 (not forcing equality). We do this sparingly – ideally just a few items. We have a rule of thumb: if $\leq 25\%$ of items per battery are freed, and when we do free them the group factor means don't shift more than ~ 0.2 SD, we consider that *close enough* to scalar invariance for practical purposes. We will, of course, report exactly which were freed.

- Alignment optimization: If many small differences exist (no single smoking gun item, but model fit is bad), we use a statistical alignment procedure (as implemented in Mplus or other tools). This procedure allows each group's intercepts/thresholds to shift a bit but imposes a penalty so they stay as equal as possible. It outputs an R^2 that indicates how well the groups' factor means can be estimated under an approximate invariance assumption. If the alignment R^2 for factor means and variances is > 0.95 , that means we can recover a common scale pretty well. In that case, we proceed with group comparisons, treating it as "approximate invariance achieved," but we will note which items had the largest biases.
- MIMIC augmentation: Another approach we sometimes do alongside is add *group predictors for each item* in a single-group model (Multiple Indicators Multiple Causes model). If group (like a dummy for each group vs. a reference) significantly predicts an item response even after controlling for F, that's a DIF item to address. This method can confirm which items need freeing or how severe bias is.

Essentially, partial invariance says: *maybe 3 out of 20 items are biased; let's account for that and still use the other 17.* Alignment says: *lots of tiny biases; let's see if overall we can still align scales without fixing each.* Both are compromise solutions that, if documented and sensitivity-tested, allow us to continue rather than declare failure.

We pre-specify that if partial invariance is adopted, we will report exactly which items were freed and even check that our main results don't change much if we drop those items entirely (sensitivity analysis). If alignment is used, we report the alignment R^2 and the distribution of the adjustments.

8.4.5 DIF (item-level fairness checks)

We dig into Differential Item Functioning (DIF) to ensure fairness item by item:

For ordinal items (IRT), a common approach is logistic regression DIF:

- We take one item at a time and fit a model with the latent trait F and group as predictors of the item response. If adding interactions or group effects significantly improves fit, that item has DIF. Specifically:
 - Uniform DIF: The group has a different intercept for that item (one group more likely to say "high" on the item even at same F).
 - Non-uniform DIF: The group has a different slope (F-item relationship differs), meaning the item's discrimination varies by group. We often use a

likelihood ratio test or compare AICs for models with vs. without group*trait interaction. Significant = flag.

For continuous items (CFA):

- We can do a multi-group LRT for each item: constrain its ν and λ_t equal vs. free and see if model fit worsens significantly when constrained – that flags DIF on that item.
- We also examine residual correlation matrices by group; if one item stands out (like consistently large residual in one group), could be item-specific DIF.

Correction for multiplicity: Because we test many items, we use e.g. Benjamini-Hochberg false discovery rate at $q \leq 0.10$ to decide which DIF findings are reliable. We don't want to overreact to one item showing a tiny p-value difference by chance. The rule is to avoid "over-freeing" – we'd rather tolerate a bit of non-invariance than erroneously free half the items because of multiple comparisons noise.

When we find DIF:

- If it's a small number of items and fits a pattern (e.g. all are certain type of wording), we might address it qualitatively (reword those items next iteration) and free them in this analysis.
- If many items show DIF, that's when alignment is a better route or we reconsider the measure in that group entirely.

8.4.6 Device and site linking (NEAT designs)

Beyond psychological differences, devices or site procedures might differ. We use linking designs:

- Common-person design: Recruit some participants who provide data on both systems – e.g. wear Device A and Device B, or go through Lab protocol and Field protocol. By analyzing their data, we can insert a linking factor in a multi-group CFA that essentially fits a linear transformation between Device A's readings and Device B's. If Device B tends to read, say, 10% higher on all HRV values, the linking factor will capture that and adjust.
- Common-item design: Ensure some measurements are identical across setups – e.g. two sites use one overlapping questionnaire or a shared calibration task. We fix those overlapping items' parameters equal across groups to act as anchors. This is like "both groups have a few test questions in common, and we use those to align the scores."

- NEAT (Nonequivalent groups with Anchor Tests): This refers to cases where the groups differ and we don't have direct overlap, but we have external anchors (like our universal anchors). For example, both Site X and Site Y do the same pain and relief tasks; we can use those anchor results as a bridge. If Site X's participants, on average, drop -0.9 HCU for cold pressor and Site Y's drop -0.7 HCU, we adjust accordingly. We might find a linear scaling factor for F such that those anchor responses line up. Then we verify after scaling that the rest of the instrument aligns (via invariance tests again).

In practice, if introducing a new device, we do a calibration study: same persons with old and new device do anchors. Then we compute a mapping (like a regression equation) from new device readings to old device's HCU. That mapping (and its uncertainty) goes into our model as part of the error budget.

8.4.7 Power and sampling considerations

All these tests require sufficient data:

- We target $n \geq 200$ per group as a rule of thumb for multi-group CFA/IRT. Smaller samples can yield unstable estimates or fail to detect moderate non-invariance. If we must work with smaller groups (say clinical subgroups that are rare), we might use Bayesian methods with informative priors to shore up power, but even then around 100 per group is a minimum for alignment methods.
- If we have many groups (like 10 languages), strict invariance tests become very sensitive (almost always will find something). In those cases, we lean on alignment with simulation-based calibration. We run simulations to see, for instance, if we had a 0.1 difference in one item's threshold, would we detect it given our sample sizes? This helps set a tolerance.
- Anchor panels per site: As mentioned, 10–20 participants per site doing anchor tasks quarterly helps stabilize the HCU transformation for that site. It's essentially a mini calibration sample to ensure that site's unit isn't drifting. We consider those anchor data in power too – if a site doesn't even have 10 people complete anchors, its HCU might be too uncertain to trust for cross-site pooling.

We plan analyses such that if invariance can't be confirmed due to low N, we default to *not* pooling. It's better to say "we don't know if it's equivalent, so we'll analyze separately" than to falsely assume equivalence.

8.4.8 Workflow (for preregistration)

Here's a step-by-step workflow that we often literally copy into our preregistration documents for multi-group scale validation:

- Specify groups: Clearly define what groups we're comparing (e.g. English vs. Spanish language, or Device A vs. B, or Age < 40 vs. \geq 40, etc.).
- Fit configural model: Allow λ , v , ψ free in each group (no equality constraints except fixing factor scale, say variance=1 in each to identify). Ensure the model structure is the same across groups. Check global fit. If poor, revise measure or drop problematic indicators before proceeding.
- Fit metric model: Constrain all factor loadings equal across groups. Compare fit to configural (Δ CFI, etc.). If Δ fit is within thresholds, accept metric invariance. If not, examine modification indices – maybe one or two loadings differ – consider partial metric invariance (though usually loadings are okay).
- Fit scalar (threshold) model: Constrain intercepts (and thresholds) equal as well. Compare fit to metric model. If fit drop is tolerable, great – scalar invariance achieved. If not, proceed to next step.
- Identify biased items (DIF/MIMIC): Use modification indices and DIF tests to find which items contribute to misfit. Free the smallest necessary set of intercepts/thresholds (partial invariance) or plan an alignment.
- Apply partial invariance or alignment: Free those specific parameters and refit, or run alignment optimization if many small biases. Document which items are freed or how alignment is done.
- Link devices/sites if relevant: If comparing instruments or sites, use NEAT linking. E.g., if raw scores differ, apply a linear transform to group B's factor to align anchor means with group A, then re-check scalar invariance on transformed data.
- Holdout validation: (Important if we care about prediction) – We might split data or have holdout channels. We ensure that after invariance adjustments, the factor can predict left-out criteria similarly in each group. For example, if F predicts “smiling” behavior, check the coefficient in each group – they should be in the same ballpark if metric invariance holds.
- Decide pool vs. stratify: Based on the results above and the rules in 8.1/8.2, decide whether it's valid to pool data or compare means across groups. If partial invariance was needed, ensure reporting reflects that (like “we compare group means after adjusting for known item biases X, Y, Z”).

By following this workflow, we maintain a transparent and consistent approach. Each preregistration includes this so that if invariance fails, we don't just quietly ignore it – we have a plan for what to do (like not testing that hypothesis across groups).

8.4.9 Reporting template (what we report)

We ensure any paper or report that uses multi-group HCI explicitly reports these invariance checks. Typically, we include:

- Model diagram or description: Showing which loadings were constrained vs. free, with group labels if needed (maybe a path diagram highlighting any group-specific parameters).
- Table of fit indices at each step: e.g. Configural fit stats, then metric, then scalar. We include ΔCFI , ΔRMSEA , etc., and note where we accepted invariance. If Bayesian, we might show Bayes Factor or posterior predictive checks.
- List of freed parameters (partial invariance): e.g. "Item 3 intercept free in Spanish" – we explicitly list them so others know the scale difference. If alignment used, we report something like "Alignment $R^2 = 0.98$ for factor means; items 5 and 12 had largest shifts" or provide a small table of item parameter differences.
- DIF results: Possibly a small table listing any item with significant DIF effect size and p-value. Or if none were significant beyond threshold, we state that.
- Device/firmware info: If applicable, we specify versions and how we linked them, plus anchor median values per group and the resulting HCU conversion factors.
- Sensitivity analyses: We demonstrate that any freed-item doesn't alter conclusions. For example, "Even if we drop Item 7 (which had bias), the difference in means remains within 0.05 HCU". Or compare strict vs. partial invariance results to show conclusions hold.

In short, anyone reading should be able to reconstruct how we ensured "same scale." Invariance isn't hidden; it's often a main table or figure in our appendices. We treat measurement invariance outcomes as results in themselves, not just preliminary analyses.

8.4.10 Failure modes and mandated responses

We also predefine what to do if invariance fails:

- Global scalar collapse: If we find that many (say > 30%) of items have different thresholds across groups and alignment R^2 is low (< .90 meaning we can't reliably put them on one scale), then we do not compare latent means at all. We would report group differences only qualitatively or not at all. Essentially, we'd treat HCI

in each group as its own scale for that analysis. We would also go back to drawing board on measure design for future.

- Metric but not scalar: If loadings are equal but intercepts aren't (common scenario), we can still compare relationships and dynamics across groups but not raw levels. So we might, for instance, compare how HCI correlates with health outcomes in each culture (that's okay under metric invariance), but we would *not* compare average HCI scores between cultures. We would state something like "Absolute HCI levels are not cross-culturally comparable due to scalar invariance failure; analyses focus on within-culture effects and associations."
- Device version effect: If a new device or firmware version changes loadings or residuals enough that invariance breaks (e.g. predictive validity drops 20%), we do not combine those data with earlier data. We either treat them as separate groups or wait until we calibrate them. We basically consider it a new instrument version (we lock versions, re-anchor, etc., as in 7.5.12).
- Channel-specific DIF cluster: If we notice, say, all GSR-related items are biased in tropical sites, that suggests maybe GSR is not reliable there. Our response: we might statistically down-weight or drop that channel and re-run invariance on the rest. Or adjust using hierarchical modeling as earlier described. But importantly, we'd document "we had to reduce reliance on channel X for those groups."

By defining these responses in advance, we maintain consistency. We won't, for example, be tempted to just ignore an invariance failure because it's inconvenient; we have a stated plan that if X happens, here's what we do (and likely that means we don't make a claim we intended to, thereby protecting from false claims).

8.4.11 Minimal code skeletons (for transparency)

To demystify this process, here are simplified code snippets (in R and Mplus) to illustrate how one would run these invariance tests:

Multi-group CFA example (using R lavaan): Suppose we have 4 items y1–y4 measuring factor F.

```
model <- 'F =~ y1 + y2 + y3 + y4'  
  
# Configural model (no equality constraints)  
  
fit_config <- cfa(model, data = df, group = "site")  
  
# Metric invariance (equal loadings)  
  
fit_metric <- cfa(model, data = df, group = "site",
```

```

group.equal = c("loadings"))

# Scalar invariance (equal loadings + intercepts)

fit_scalar <- cfa(model, data = df, group = "site",

                    group.equal = c("loadings", "intercepts"))# Compare metric vs. scalar

anova(fit_metric, fit_scalar) # or use fitMeasures to get ΔCFI, ΔRMSEA

```

This would yield whether intercept constraints significantly worsen fit. We'd inspect modification indices (modindices(fit_scalar)) to find which intercepts are causing trouble.

Graded IRT with DIF (using R mirt): For ordinal data:

```

# Configural (no invariance)

mod_config <- multipleGroup(data_ord, model = 1, itemtype = "graded", group = "lang")

# Metric invariance (equal slopes 'a')

mod_metric <- multipleGroup(data_ord, model = 1, itemtype = "graded",

                             group = "lang", invariance = c("slopes"))

# Scalar invariance (equal slopes and intercepts 'd')

mod_scalar <- multipleGroup(data_ord, model = 1, itemtype = "graded",

                             group = "lang", invariance = c("slopes", "intercepts"))

# DIF test on mod_metric for item parameters 'a1' (slope) and 'd' (intercepts)

DIF(mod_metric, which.par = c("a1", "d"), scheme = "add")

```

The DIF function would flag items with significant slope or intercept differences.

Alignment in Mplus (syntax snippet): If we had 8 items y1–y8 across 4 language groups:

ANALYSIS: ALIGNMENT = FIXED; ESTIMATOR = MLR;

MODEL:

F BY y1-y8;

MODEL GROUPS: language(4);

OUTPUT: ALIGNMENT;

This would output estimated group means and variances under alignment, and flag non-invariant items. We'd look at the "alignment output" section for R^2 etc. These code snippets aren't comprehensive but show the gist of how one actually tests invariance.

8.4.12 Ethics and governance notes (for measurement across groups)

Finally, we acknowledge some higher-level points:

- Transparency in methods: We treat invariance testing results as part of the scientific findings, not just an appendix. If a scale doesn't hold up in one group, that is information. We publish those diagnostics alongside the main results, not bury them. This way, if later someone tries the scale in a new context, they can anticipate issues. Also, if HCI "fails" somewhere, it's on record.
- Equity in measurement: If one group doesn't fit the scale, we don't force it just to have them "included." For example, if scalar invariance fails in a certain cultural minority, we won't simply fix the parameters equal and go on to claim "no difference." That would effectively erase real differences and could lead to conclusions that ignore that group's reality. Instead, we highlight that our instrument is not (yet) fair for that group. In practice, that might mean developing a tailored version or improving translation.
- Versioning of instruments: We consider each major recalibration a new version of HCI/HCU. If we adjust anchors or loadings, that's a version update, which we document similar to software. Any time we present results, we state "using HCI 2.1" so that others know exactly what definition was used. That way, if someone else is comparing or meta-analyzing, they'll account for version differences.

Takeaway: The ladder of Configural → Metric → Scalar invariance is how we make "same scale" more than just a hope – it becomes a tested achievement. By climbing it carefully (allowing partial invariance, using alignment, linking devices) and being honest about the results, we earn the right to say "Person A's HCU = Person B's HCU". And crucially, when the ladder breaks (invariance fails), we stop climbing and report that limitation. In essence, we don't just assume everyone's data is on the same ruler – we prove it, or we don't do the cross-group comparison. This discipline is what underpins all our pooled analyses in testing the Law of Fairness.

8.4.13 Where we go next:

With formal tests in place, we turn from "can we compare?" to "how do we build comparability step by step?" 8.5 presents the calibration ladder—within-person, within-site, then cross-site—so teams can climb to credible generalization without skipping rungs.

8.5 Calibration Ladder (Within → Between → Cross-Cultural)

You cannot jump straight to “global comparability” in one leap. Instead, HCl/HCU must earn transportability one rung at a time. We envision a ladder of calibration with three main rungs:

1. Within-person calibration: Make sure the instrument is stable and sensitive for each individual over time and context.
2. Between-person (within a site) calibration: Ensure the scale works comparably among different people in the same environment or study site (so you can aggregate individuals within that group meaningfully).
3. Cross-site/cross-cultural calibration: Finally, extend the scale across different labs, cultures, or populations, once the first two levels are solid.

Each rung has its own tools, criteria, and “fail-safes.” We only climb to the next rung if the previous one is sound. This staged approach prevents us from, say, declaring a universal scale works everywhere when we haven’t even checked it works reliably within one person or one lab.

8.5.1 Rung 1 — within-person calibration (stability and sensitivity)

Goal: Demonstrate that HCl can track the *same person’s* affect consistently and meaningfully across time, devices, and contexts. Essentially, within one individual: if their actual feelings change, HCl should change; if they are in a stable period, HCl should be stable. Also, define what a ±1 HCU change means *for that individual* week to week.

Design: We typically do an intensive longitudinal study per person:

- Duration: e.g. 2–4 weeks of continuous or daily monitoring for each participant.
- Channels: Use all core channels (self-report EMA, HRV, pupil via wearable, maybe SCL if available). We might not include brain measures here for practicality, except maybe one overnight EEG subset.
- Anchors: We include at least a couple of micro-anchor sessions for each person (like a weekly cold pressor or a relief event). And possibly one lab session with a full anchor (pain or chills) about mid-way.
- Device swap: Mid-study, we might intentionally swap out their device (e.g. have them wear a different sensor for a few days) or have them wear two in overlap, to test device linking within person. This is critical to see if device differences can be bridged at individual level.

Analyses (Within-person metrics):

- Test-retest reliability: Compute an intra-class correlation (ICC) across adjacent weeks for baseline HCI level. We aim for ICC (2,1) around ≥ 0.70 for weekly averages. This shows that when a person is in a similar state week to week, HCI yields similar values (i.e., not pure noise).
- Convergent anchor responses: Check that the *same person's HCU response to an anchor is repeatable*. For instance, if in week 1 a cold pressor caused -0.5 HCU for them, in week 3 a similar cold pressor should cause roughly -0.5 HCU (within say ± 0.3). Consistency here means the scale is calibrated for that person (they're not randomly giving -0.2 one time and -1.0 another for the same stimulus unless something about them changed).
- Device agreement: If two devices or methods measured the person concurrently, do they agree on HCI? Use Bland–Altman limits of agreement to see differences. We want, for example, that the 95% agreement is within ± 0.4 HCU and no systematic bias between devices.
- State–trait variance: Use a hierarchical model to decompose how much of the variance in HCI is at person level vs. day level vs. moment level. If we find that, say, 30% of variance is between-person (trait-like) and the rest is within-person, that's good – it means within-person fluctuations (state) are a big part but there's also a stable individual baseline. We'd expect some stable component but also dynamic range.
- Possibly also correlation with other within-person measures: e.g., does their daily HCI correlate with daily self-reported mood in expected ways, etc., just to ensure it's capturing something real.

Decision rule (to move to Rung 2): We only proceed to comparing between different people if all within-person criteria are met:

- ICC (week-to-week) ≥ 0.70 ,
- Anchor responses consistent within ± 0.3 HCU,
- Device agreement acceptable (no major bias, LOA within threshold),
- State–trait structure showing a reasonable portion of variance at person level (we don't specify exact %, but we want to see a stable component, meaning the instrument isn't pure noise).

If these fail, we *stop*. For example, if test–retest is low (HCI all over the place for each person), then adding more people won't help – the instrument itself is not reliable. We would then remedy before proceeding: check for device calibration issues, re-run anchors, see if a channel was malfunctioning, etc.

Failure actions (if Rung 1 fails):

- If ICC is low or drift present, we investigate if devices were inconsistent or if anchor schedule wasn't enough, and recalibrate, possibly extending within-person tracking until stable.
- If devices show a bias, we calculate a linking function (like a regression between Device A and B readings) and decide either to adjust data or to not mix those devices' data going forward.
- If anchors were unstable (maybe participant didn't follow instructions one time, etc.), we might repeat anchor sessions or exclude that participant from between-person comparisons (maybe some individuals just didn't engage properly).

We only feel confident comparing people to each other if each person's measure makes sense internally.

8.5.2 Rung 2 — between-person, within-site (comparability among peers)

Goal: Now assume each person's HCl is consistent for them. Next, show that people measured under the same conditions (same team, same city, same device type) can be compared to each other fairly on their HCU levels and responses. Essentially, calibrate a common scale within a homogeneous group. This tests whether, for example, one person's 6/10 happiness is similar to another's 6/10 in that context, and whether differences in trait or reactions are real and measurable.

Design: Usually a standard study sample from one site:

- Sample: Aim for $N \geq 150$ from one population (like one lab's study, possibly diverse in gender/age, maybe intentionally include subgroups like "chronic pain patients" vs. "healthy" to test known differences).
- Anchors: Everyone goes through the pain and chills or exclusion tasks at least once so we have cross-person anchor data. Also, we run quarterly (or at study start) anchor panels to ensure no drift as data collection spans time.
- Tasks: We include tasks designed to elicit Queue System signatures: e.g. a decision-making task with varying horizons, or a scenario where they choose between an immediate reward vs. altruistic act (to see QS effects). The exact tasks aren't for calibration per se, but their results will be used to test predictive validity of HCl in this sample.

Analyses (Between-person metrics):

- Measurement invariance across subgroups: If our sample includes notable subgroups (like an equal split of men and women, or younger vs. older, or patients

vs. controls), we test configural/metric/scalar invariance between those subgroups. This is basically repeating Chapter 8.4's steps but within this single site, to confirm HCI works the same for sub-populations. We might find we need partial invariance (e.g. one questionnaire item works slightly differently for older participants – we'd adjust).

- Anchor alignment: Check that this site's anchor medians align with the reference definition. For instance, if globally we said $-1 \text{ HCU} = \text{median drop in cold pressor}$, does this site's median drop equal -1 HCU or close? If not, maybe this site's anchors are a bit off; we either document that (like “our participants had slightly lower pain sensitivity, so their median drop was -0.8 HCU ”) or we adjust scale slightly. But ideally, it matches within tolerance (say ± 0.3).
- Predictive validity: A big one – we test whether adding F_t (our HCI latent) to certain outcome models improves prediction as expected. For example, does HCI predict who will choose to abort a task early (maybe those with persistently low HCI quit sooner)? Or does it correlate with known neural signals (vmPFC activation)? We use preregistered models: e.g. a logistic regression for a choice, with and without HCI as a predictor. If adding HCI significantly improves out-of-sample prediction (or model fit), that means HCI is capturing something behaviorally or biologically relevant (convergent validity).
- Known-groups validity: If we deliberately included, say, a group of chronic pain patients expected to have lower baseline happiness, we check that indeed their HCI is lower and that their response to analgesic anchor is larger than controls, per hypothesis. These are like “positive controls” – if HCI fails to pick up a difference we expect, that's concerning.
- Spec-curve robustness: We might run a multiverse of reasonable data processing pipelines (slight variations in filtering, scoring, etc.) to see if the central results (like mean differences or correlations) hold median positive. If HCI's conclusions flip under slight analysis changes, that would indicate fragility.

Decision rule (to move to Rung 3): We only proceed to generalize across sites if key conditions are met in this sample:

- (If relevant) Scalar or at least partial invariance holds within this site's subgroups (so we're not dealing with undetected heterogeneity).
- Anchors for this site are properly aligned (no big unexplained deviations).
- HCI shows predictive validity: it adds explanatory power to known outcomes (our QS markers, etc.). For example, if adding HCI to models yields ~ 0 improvement, then HCI might be capturing nothing new – a fail condition (we had one in Chapter 7.6: if HCI doesn't predict beyond standard measures, it's questionable).

- Pre-registered effects appear: e.g. QS predicted that as horizon shrinks, HCl effect on decisions intensifies; if we saw nothing of the sort, either QS is wrong or HCl isn't measuring properly. If all such tests are null, we reconsider instrument validity (maybe calibrate differently or acknowledge limitation).

If these pass, we are confident that within this environment, we have a working common scale.

Failure actions (if Rung 2 falters):

- If scalar invariance didn't hold, we do partial invariance or alignment as in 8.4 and then do not compare means between those subgroups beyond what's supported. E.g. "We can compare patient vs. control reactivity, but not their absolute HCl levels because item X biased."
- If adding HCl yields no predictive gain, that's serious. We would report that result (that HCl didn't help in e.g. predicting choices) and then likely not attempt a law-level test with it (maybe we'd refine the index first). It doesn't necessarily break the scale, but it fails a reason for the scale's existence. We might revisit nuisance modeling: maybe arousal wasn't properly accounted, overshadowing HCl.
- If anchors drifted (say by the end of the study, re-testing anchor shows people responding differently), we recalibrate or at least handle that (split data before vs. after drift, etc.) and certainly would pause before multi-site aggregation.

8.5.3 Rung 3 — cross-site / cross-cultural transport (the real test)

Goal: Finally, show that HCl/HCU can travel across different languages, cultures, climates, devices and still be comparable. This is the ultimate test: that our measurement of affect has a stable meaning globally (or identify exactly where it doesn't). If this succeeds, we can truly pool data for large-scale LoF tests; if not, we delineate the limits (e.g. maybe the scale only works within Western cultures for now, etc.).

Design: A dedicated multi-site study:

- Sites: At least 3 distinct locations (e.g. different countries or at least different cultural groups), using at least 2 different native languages among them. Also, use at least 2 different device hardware stacks if possible (to test tech portability).
- Each site runs the full anchor triad (pain, chills, exclusion) and the QS tasks (like in Rung 2) with local participants.
- Bridge samples: Include special samples to help linking: e.g. a small bilingual group that does the tasks in both languages; or one site where we use both types

of devices on same people (common-person design for devices). Also ensure some items or procedures are common across all for anchor (we did that by design, all have same anchor tasks).

- Context covariates: We record things like climate (temp/humidity) for each site during testing, altitude, etc., and also differences like typical sleep patterns or diet if relevant (these might be nuisance in analysis).
- Possibly involve an exchange: one site's team travels to another to run the study in exactly the same way to see if differences are procedure or sample (if resources allow).

Analyses (Cross-site metrics):

- Invariance ladder across sites: Now we do multi-group invariance with *groups = sites* (and possibly *groups = languages* separately). We test configural (does HCI factor structure hold everywhere), metric, scalar invariance across these sites. Partial invariance or alignment likely needed if any cultural biases in items; we apply those as per 8.4. The criterion might be alignment $R^2 \geq .95$ for means – meaning we can estimate a common scale well.
- NEAT linking of anchors: Use the anchor outcomes as “equators.” For each site, compute the average HCU change for each anchor and see how it compares. If one site's cold pressor gave only -0.5 HCU while another's gave -1.0 , that suggests either population pain tolerance difference or slightly different execution. We then apply a linear transform to that site's HCU scores (e.g. multiply by some factor) to align those anchor medians. We then re-check invariance on the transformed data to see if it improved.
- Transport diagnostics: Compare key model outputs across sites: for example, horizon scaling slopes (does shortening time horizon increase HCI's effect on decisions similarly in all sites?). We require at least that the direction of effects is consistent and magnitudes within a pre-registered band (maybe $\pm 20\%$ of each other). If one site showed a completely opposite result (say, horizon had no effect or reverse), that could mean either theory fails or measurement failed there.
- Cross-site generalization: We do something like train a predictive model on data from Sites A+B and test on Site C. If our HCI truly means the same, a model built on combined data of some sites should still predict behavior in a new site pretty well (maybe within 70% of original performance). This is a stringent test: it checks both measurement and whether underlying behavioral relationships hold.
- Error budgets site-wise: We calculate how much of variance in outcomes or in HCI itself is attributable to site differences. If one site accounts for $>40\%$ of variance (meaning site differences dominate person differences), then maybe pooling is

not safe. Ideally, site only contributes modestly after accounting for all our linking adjustments.

Decision rule (pool vs. stratify globally): We only pool data across sites for law-level tests if:

- Metric invariance holds across sites and either scalar invariance holds or alignment indicates factor means are comparable ($R^2 \geq .90$ and $\leq 25\%$ items freed as a guideline).
- HCU transforms from anchors have converged (no site's anchors consistently outside ± 0.3 HCU range of pool).
- Cross-site predictive patterns (like QS signatures) replicated within tolerance (we set tolerances in preregistration).

If those are not all met, we will not lump sites together for a single global analysis. We would either stratify (analyze each culture separately or use it as a moderator) or explicitly include site as a factor.

We also declare the scope clearly: e.g. "We pooled data from US, Iceland, and Korea, but kept data from Riyadh separate due to residual non-invariance in SCL channel". So LoF claims would then apply to that pooled set, but not necessarily to the excluded site until fixed.

Failure actions (if Rung 3 partially fails):

- If a particular site diverges (e.g. one country's data just doesn't fit well on our scale), we might *down-weight certain channels* for that site (like earlier example, down-weight SCL in tropics) and try alignment again. Or refine translations and retest later.
- If still problematic, we treat that site as local-only: we would report their results separately and not mix them. This is not a failure of LoF necessarily, but a scope limitation: "the instrument is not yet validated in X context, so we can't test LoF there until improvements are made."
- We'd publish all diagnostics so that it's clear where the scale works and where it doesn't.

8.5.4 The ladder as a pipeline (operational checklist)

To operationalize the above, our team essentially runs a pipeline moving through these rungs. For clarity, here's what that pipeline looks like in practice:

Within-person phase (Rung 1):

- Recruit ~30–50 individuals for intense monitoring over a few weeks.
- Have them do weekly micro-anchors and at least one lab anchor mid-way.
- If multiple devices, include a swap or overlap period.
- Compute reliability (ICCs), anchor repeatability, device LOA.
- If all checks out (criteria met), *green light* to next phase. If not, recalibrate or redesign and possibly repeat within-person until stable.

Within-site (between-person) phase (Rung 2):

- Enroll larger sample (150–250) at one site, potentially including some known groups.
- Everyone runs through anchor triad and QS tasks.
- Perform multi-group (if subgroups) invariance, predictive validity analysis.
- Gate: Only if invariance acceptable and HCI proves useful (predictive, differentiates known groups) do we move on.
- If criteria not met, address issues (free biased items, adjust index formula, etc.) and possibly iterate within that site.

Cross-site phase (Rung 3):

- Initiate 3–5 sites, ensuring diversity in language/culture and including bridging elements.
- All sites implement anchors and tasks as uniformly as possible; collect bridging sample data.
- Run linking (NEAT), multi-group invariance across sites, compare results.
- Decide whether to pool most sites or stratify if needed; set any site-specific adjustments (like hierarchical model weights) and uncertainties.

Maintenance ongoing:

- After establishing, continue quarterly anchor panels at each site to detect drift.
- Log any firmware/software changes; if any big changes, re-run a mini Rung 1 at that site to ensure within-person still good.
- Annually, re-assess invariance and calibrations with accumulated data; publish an updated HCU “version note” if anything changes (e.g. “we adjusted anchor definition slightly based on new data”).

By formalizing this pipeline, whenever a new lab or context wants to join the “network,” we essentially walk them up these rungs. We don’t just plug in data from a new context into a global analysis without individually validating them through Rungs 1 and 2 first.

8.5.5 Worked illustration (fictional but realistic)

To make the ladder concept concrete, consider this scenario (numbers invented but plausible):

- Rung 1 (Boston site, internal test): We follow 40 individuals intensively. Result: weekly HCl baseline ICC = 0.78 (good). Pain relief anchor (nitrous oxide) raised HCl by +1.15 HCU in week 2 and +1.10 in week 4 on average – repeatable within 0.05. Wearing a smartwatch vs. a chest HR strap gave HCl readings that differ by at most ± 0.32 HCU (no bias, just noise) – acceptable. Pass Rung 1.
- Rung 2 (Full study in Boston): 180 participants. Multi-group invariance by gender: needed to free 2 item thresholds for scalar invariance (perhaps men rated one item differently), but partial invariance achieved. HCl improved choice prediction AUC by +0.06 over base models (significant). A key QS effect (horizon \times HCl interaction) was robustly positive as expected. Spec-curve of analysis variations all showed a positive effect. Pass Rung 2.
- Rung 3 (Three international sites – Reykjavík, Riyadh, Seoul): Data comes in. Metric invariance holds; scalar handled via alignment with $R^2 = .97$ (almost full invariance). We did notice SCL channel had to be down-weighted in Riyadh (due to climate sweat) – we adapt model for that. Anchor medians: all within 0.28 HCU of the Boston reference (pain maybe a bit lower in Icelanders, but within tolerance). QS-residual patterns (the effect of prior frustration on HCl) replicated in all three new sites. We decide to pool Iceland, US, Korea. However, in Riyadh, even after down-weighting SCL, we weren't fully confident (some non-invariance remained in a few items), so we choose to stratify Riyadh's data – i.e. include it but analyze it separately or with group as a moderator. We publish channel weight differences and error budgets for transparency (e.g. “in Riyadh, HRV and pupil contributed 70% of HCl weight, vs. 50% in other sites, due to reduced SCL weight”).

This fictional scenario shows how at each rung we made decisions and adjustments, but overall succeeded in building a common scale across multiple contexts (with one partial exception which we handle via stratification).

8.5.6 Statistical guardrails and preregistration snippets

We predefine some statistical guardrails to keep ourselves honest:

- Stopping rules: If we don't meet the criteria of a rung, we do not proceed to the next rung (we even mention this in preregistration: e.g. “If after Phase 1 $ICC < 0.7$, we will not aggregate across participants”). We would instead pause and

redesign. We even have a term “ladder stop” to declare in a report if we had to stop – meaning “we could not validate cross-person comparison, so we didn’t do it.”

- Equivalence tests for anchors: We use TOST (Two One-Sided Tests) on site anchor medians to statistically confirm they are within ± 0.3 HCU of each other. Essentially, test the null that difference is bigger than 0.3 and want to reject that. This provides evidence the units are practically equivalent across sites.
- Non-inferiority margins for prediction: For cross-site predictive validation, we set a margin like “new site’s predictive performance must be no worse than 20% below the original site’s performance”. We then statistically test if performance drop is within that margin (non-inferiority test). This ensures we don’t declare success if, say, the model only works in original site and flops elsewhere.
- Multiplicity control in cross-cultural tests: We plan our site comparisons such that if we compare, say, 3 sites pairwise on anchor AUCs or something, we apply Bonferroni or Holm correction. We’d rather be cautious than claim something universal that’s actually an artifact.

By including these in analysis plans, we avoid cherry-picking positive results – we set what success looks like and what failure triggers.

8.5.7 Governance: versioning the scale

As mentioned, HCI/HCU is not static; we treat it like software with versions:

- We assign a version ID to each major release of HCI/HCU (e.g. HCU Version 1.0 for initial, then 1.1 if we tweak something) and stamp that on every dataset and figure we produce. This way, if data collected under an older calibration is used, it’s labeled accordingly.
- We maintain change logs: if we change anchor procedures (say in 2024 we switch from cold pressor to a different pain inducer because it’s safer), or if device firmware updates happened, or if we tuned a hyperparameter in the state-space model, all these are logged in a document that accompanies publications.
- We provide backward compatibility info: for example, if Version 2.0 of HCU scales pain differently, we publish a conversion function or at least say “Multiply old HCU by 1.1 to get new units, with ± 0.1 margin of error”. We also inflate uncertainty appropriately when converting old to new (because conversion isn’t exact).

This governance ensures that when we say results across studies, we know they’re using the same measurement definition or we adjust for differences. It’s akin to how IQ tests are periodically renormed and one must specify which edition was used.

8.5.8 When to retreat (and how to say it)

We also define conditions under which we would retreat on claiming a global scale:

- If even after remediation, within-person repeatability fails (say we just can't get consistency), we would not proceed to use that measure for group comparisons. We'd likely publish a methods note: "HCl did not achieve required reliability, therefore we cannot use it for between-person tests yet".
- If within-site scalar invariance is impossible (e.g. > 40% items biased across subgroups), we would conclude that HCl can only be used for *within-person or within-group changes* in that context, not for comparing levels. Essentially we'd downgrade it to a relative measure only.
- If at cross-site level alignment $R^2 < .90$ or if cross-site effects actually conflict (sign flips, etc.), we would not pool globally and explicitly state "external validity across these cultures was not established." We would then either stratify by culture in all analyses or declare that global generalization is beyond current evidence.

In all cases, we *say it clearly*. We won't quietly ignore that, for instance, Asia data didn't fit – we'd write something like "The HCl could not be reliably transported to X context; thus, results for X are presented separately and not considered part of the unified support for LoF."

Takeaway: The calibration ladder turns the lofty goal of "a global fairness metric" into a sequence of earned steps. We confirm it works *within individuals* first, then *within a homogeneous group*, and only then *across different groups*. At each step we have thresholds and transparency. By climbing carefully and stopping when needed, we ensure that if we do reach the top (global scale), it's on solid ground. This approach guards against over-claiming universality and helps pinpoint where things need improvement. In the end, if we do succeed, we have numbers that deserve to travel – and if not, we know exactly why and in what way they fell short.

8.5.9 Where we go next:

The ladder gives us a path to comparability; the ledger still needs honest error bars. 8.6 shows how uncertainty propagates from indicators to latent state to lifetime totals, and how decisions about neutrality should be reported as probabilities, not slogans.

8.6 Propagating Uncertainty into the Ledger

A life's ledger of well-being (the integral of F over time) is not a single precise number – it's an uncertain estimate. Any claim like “this person's life had zero net suffering” must acknowledge error bars. Science demands we quantify that uncertainty at every step: from sensor noise, to model assumptions, to missing data, to unit conversions. Chapter 8.6 details how we carry forward all these uncertainties right into the final ledger, and how we visualize and make decisions with them. In short, a ledger that pretends to be exact isn't credible. We instead produce a distribution or confidence interval for the ledger total, and use that to judge neutrality or fairness claims probabilistically.

8.6.1 Sources of uncertainty (error budget recap)

We maintain an error budget – an accounting of all major uncertainty sources in HCI/HCU. Key components include:

- Measurement noise: Within each channel, readings have noise. E.g., heart rate sensors have jitter, self-report has random mood fluctuations or momentary errors, EDA sensors might drop out. Different devices add variance. Also includes missing data – e.g., if 10% of data is missing, that adds uncertainty (even if imputed). This is usually quantified by sensor error variances and any data imputation uncertainty.
- Mapping error: Uncertainty in how observed signals map to the latent F . In our factor model, this is reflected in standard errors of loadings and intercepts, especially if we allowed some group-specific variations. If partial invariance was used, there's error about those adjustments. Essentially, how well we estimated the true relationship between each channel and F .
- Model (process) error: Our state-space model of how F evolves over time has assumptions (e.g. that F follows a local linear trend with certain process noise q). There's uncertainty in those parameters (like q , or seasonality effects). If the model is slightly mis-specified (maybe true emotional dynamics are more complex), that introduces error in F estimates. We capture this as process noise variance and perhaps compare multiple models (process noise can be learned via Kalman filter residuals etc.).
- Anchor/scale error: We define ± 1 HCU by finite sample anchors. If our reference group was small, the anchor effect size has a confidence interval. Also, each site's local anchor calibration has error (when we say “Site B's values were multiplied by 1.1 to align,” that 1.1 has an uncertainty). This contributes to uncertainty in any absolute HCU values.

- Linking error: If we link devices or languages through statistical adjustments (NEAT, alignment), that has error. E.g., we might say “add 2 points to all scores from device X,” but that 2 has ± 0.5 margin. Or alignment yields group means with some standard error.

We formalize \mathcal{E} = total error budget, which is essentially the variance/covariance of final HCU or ledger estimates accounted from all these pieces. The goal is that any number we present (like ledger total) comes with a confidence interval derived from this \mathcal{E} .

8.6.2 From indicators to latent F with variance

Let’s break down how we propagate uncertainty *at the momentary level first* (before integrating over time). We estimate the latent affect F_t at each time with a hierarchical state-space model, which naturally gives us a probability distribution (not just a point) for F_t given the data.

Observation layer: For each channel k , we had equations:

$$y_t^{(k)} = v_k + \lambda_k F_t + \Gamma_k Z_t + \varepsilon_t^{(k)},$$

with $\varepsilon_t^{(k)} \sim \mathcal{N}(0, \Psi_k)$ (if continuous).

For ordinal indicators:

$$P(y_t^{(k)} \geq c) = \text{logit}^{-1}[a_k(F_t - b_{-k}c)].$$

Each of these equations has *parameters with uncertainty* (we have posteriors or confidence intervals for λ_k , etc.) and each observation yields some likelihood of F_t .

State layer: We model F_t dynamics often as:

$$F_t = \mu_t + \eta_t, \quad \eta_t \sim \mathcal{N}(0, q) \quad (\text{observation noise floor})$$

$$\mu_t = \mu_{t-1} + \delta_{t-1} + \zeta_t, \quad \zeta_t \sim \mathcal{N}(0, \sigma_\mu^2) \quad (\text{level/trend})$$

$$\delta_t = \delta_{t-1} + \xi_t, \quad \xi_t \sim \mathcal{N}(0, \sigma_\delta^2). \quad (\text{slope random walk})$$

This is a local linear trend model; we may add cyclic terms for daily rhythms or horizon shrinkage. Importantly, it has process variances $q, \sigma_\mu^2, \sigma_\delta^2$ that we fit or set, which reflect how unpredictable F is over time (i.e., fundamental uncertainty in the mind’s evolution).

We use filtering (Kalman filter if we linearize ordinal parts, or particle filter if needed for heavy nonlinearity) to compute the posterior distribution of F_t given all data up to time T . The result is not just \hat{F}_t but also $V[F_t]$ the variance of the estimate.

In practice:

- If all channels are Gaussian, Kalman filter gives mean and covariance recursively.
- If not, we might do an extended Kalman or unscented or particle filter to approximate $p(F_t | y_{1:t})$.
- Either way, we get an estimate with an uncertainty (like $F_t = 0.5$ with SD 0.2 at time t).

These posteriors incorporate observation noise (ε) and model process noise (ζ, ξ) and they also reflect missing data (if at a time we had no data, $V[F_t]$ will blow up based on the prior). Importantly, this method retains correlations over time – if we have a gap, uncertainty grows and that correlates with uncertainty in the next times.

So by the end, for each time point or each segment, we have F_t distributions.

8.6.3 From latent draws to ledger draws

To propagate uncertainty into the ledger, we sample latent trajectories F and map each draw into the momentary HCl scale before integrating over time:

We take many draws (simulations) from the posterior distribution of the entire F trajectory over the period of interest. For example, suppose we monitored someone for T hours/days, we have $p(F_0:T | \text{data})$ – typically we can sample from our state-space model's posterior (e.g. via a forward filtering backward sampling algorithm for Kalman filters).

For $m = 1$ to M (like $M=1000$ samples):

- Sample an entire trajectory $F_{0:T}^{(m)}$ from the posterior.
- Numerically integrate the mapped trajectory: $L^{(m)}(t_j) = \sum_{i \leq j} HCl^{(m)}(T_i) \Delta T_i$ (like a Riemann sum). If data is regular, it simplifies to sum; if irregular, we naturally handle varying Δt as indicated. We might use trapezoidal rule or Simpson's if needed for better accuracy over bigger intervals, but often high frequency is enough that simple sum is fine.

This yields a set $\{L^{(m)}(T)\}_{m=1}^M$ – a distribution of ledger totals at time T . We then define, for instance, the 95% credible interval as [2.5th percentile, 97.5th percentile] of these draws $L^{(m)}(T)$. That is our uncertainty band for the ledger at final time.

If computationally heavy, one can approximate by linearizing: There's a delta-method formula: $V[\int_a^b F_t dt] \approx \int_a^b \int_a^b Cov(F_u, F_v) du dv$. This double integral of covariance basically sums up how uncertainty in F accumulates over the integration window – accounts for the fact uncertainties might cancel or amplify depending on correlation of F at different times. In a state-space with known covariance structure, one can get that (like integrating

the process noise over time with correlation). But that formula is complicated and we prefer simulation because it naturally handles non-linearities and irregular sampling. We mention it as an approximation for short windows if needed.

The key: at the end of day or end of life, we don't report "L = 5 HCU" – we report something like "L = 5 ± 1.5 HCU (95% CI [2, 8])". And if we suspect any biases, we include them such that 0 might or might not be in that interval.

8.6.4 Injecting anchor and linking uncertainty

Up to this point, we treated the conversion from latent affect F_t to the observed HCl scale as fixed. More precisely, we map latent affect F_t into the momentary HCl(t) scale (in HCU per unit time) using anchor-defined scaling. Formally,

$$\text{HCl}(t) = aF_t + b,$$

where $a > 0$ is a scale parameter preserving directionality of affect and b is an offset, both determined by calibration anchors. Previously we treated a and b as fixed constants. In reality, they are estimated quantities and therefore carry uncertainty.

To propagate this correctly, we jointly sample a and b from their posterior distribution along with the latent trajectories $F_{0:T}$. Importantly, a and b are drawn jointly to preserve any posterior covariance between scale and offset.

We estimate, from anchor panel data across sites, a joint posterior distribution for (a, b) . For example, a might be centered near 1 HCU per unit of F with some uncertainty, while b is ideally near 0 but may deviate slightly if alignment or device linking requires a small offset.

For each Monte Carlo draw m , we sample:

$$(a^{(m)}, b^{(m)}) \sim p(a, b \mid \text{anchor data})$$

alongside a sampled latent trajectory $F^{(m)}(t)$.

We then transform the latent trajectory into HCl units:

$$\text{HCl}^{(m)}(t) = a^{(m)}F^{(m)}(t) + b^{(m)}.$$

Each trajectory draw is therefore expressed on the anchored HCl scale, already incorporating scale and offset uncertainty. When we integrate these transformed trajectories to compute ledger draws $L^{(m)}(T)$, the resulting distribution inherits calibration uncertainty in addition to state-estimation uncertainty.

We apply the same principle to any device or site linking parameters. In each Monte Carlo draw, we sample linking parameters from their posterior distributions and apply the corresponding transformation before integration. Linking uncertainty is thus propagated directly into the final ledger intervals rather than treated as fixed.

This procedure typically widens ledger intervals slightly, especially over long integration windows, because small calibration differences accumulate over time. For example, uncertainty in the offset b induces a linear contribution to integrated uncertainty proportional to total duration. By explicitly propagating these terms, we avoid understating confidence in long-run ledger estimates.

In short, the ledger is not computed on a single fixed scale; it is computed across a distribution of plausible calibrated scales. All reported neutrality probabilities and interval estimates therefore reflect both measurement and calibration uncertainty.

The outcome is a more honest assessment: e.g. “We estimate 0 net HCU with $\pm X$, but if our calibration was 5% off, that would mean an additional $\pm Y$ uncertainty.” It’s all baked in.

8.6.5 Daily and weekly summaries with intervals

It’s often useful to look at summaries like “how was this day in HCU?” or “the last week’s total.” We propagate uncertainty there too:

- Daily integral: $L_{\text{day}} = \int_{00:00}^{24:00} HCl(t) dt$ for each day, with a 95% CI. We get this from the trajectory samples easily by restricting the sum to that day. So we can say “Tuesday: 2.3 HCU ± 0.5 ” (units are HCU \times time, which we might call HCU-days if integrated over a day).
- Rolling weekly (7-day) sum: because weeks overlap (like Monday–Sunday, then next Monday–Sunday overlaps 6 days), we must account for covariance in overlapping windows when summarizing uncertainty. We handle it by either explicitly computing overlapping window distributions from draws or by not treating adjacent weekly points as independent in visualization.
- Counterweight index: We define metrics like the ratio of positive to negative area after big events (see Chapter 6). We get that for each draw too and hence a distribution. For example, after a painful event, maybe the counterweight index is 1.2 with a 95% CI of [0.8, 1.6] from bootstrap or Monte Carlo (where we sample several realizations of recovery segment).
- MDC (Minimal Detectable Change): Based on the error budget, we compute e.g. the smallest 24h HCU change that would be statistically significant (usually something like $1.96 * \text{SD}$ of 24h measurement error). If we find, say, $MDC_{24h} =$

0.35 HCU-day, that means any daily difference less than that is likely noise. We report these MDC values so that if someone's ledger changes by 0.2 from one week to next, we can say "not above noise level."

All these summary stats come with uncertainty. We emphasize that in any public or clinical use, one should interpret, say, "yesterday was +0.5 HCU-day" with the knowledge of \pm range. Even for individuals tracking their own ledger, we'd show them a band ("you likely had between +0.2 and +0.8 today; we're 95% sure it was positive").

8.6.6 Probability of neutral closure and decision rules

In the end, if someone asks: did this person's life (or year) end in the negative, positive, or neutral? We answer with a probability, not a yes/no.

Define "neutral" band $[-K, K]$ in HCU. Then: $p_{\text{neutral}}(T) = P(L(T) \in [-K, K] \mid \text{data}) \approx \frac{1}{M} \sum_{m=1}^M I\{L^{(m)}(T) \in [-K, K]\}$.

We just see what fraction of our ledger draws ended between $-K$ and $+K$. For instance, if $K = 0$ (exact neutrality), we might get $p_{\text{neutral}} = 0.7$, meaning 70% of sampled trajectories had net ~ 0 (some positive, some negative swings ended up balancing). If $K > 0$ to allow a margin, that probability increases.

We would report that: e.g. "Probability the ledger is neutral (within $\pm K$ HCU) at $T = 10$ years is 85%".

Our decision rule might be:

- We don't declare "neutrality achieved" unless p_{neutral} is very high (≥ 0.95) and we see no trend in the last segment. We require not just probability, but also that the median ledger has flattened – if it's drifting up or down even within the interval, we wouldn't be comfortable calling it stable yet. For example, if last 7 days show monotonic change, maybe the compensation process isn't done, so we wouldn't say neutral even if currently the mean is near 0 (because it might overshoot).
- If p_{neutral} is less, say 0.5, it's very inconclusive; if it's very low (< 0.05), we'd be pretty sure it's not neutral.

For comparing groups' ledgers (like two people or two policies), we similarly wouldn't just compare point estimates. We'd use equivalence testing on distributions: e.g. use a TOST to see if difference in L distributions is within a small bound. If yes, then declare them effectively equal; if not, see which dominates.

The overarching idea: neutrality or fairness outcomes become statements of probability. We don't say "Yes, the ledger balanced; LoF satisfied." We say "Given all data, there's a

93% chance this ledger balanced to within ± 5 HCU-days.” And we set a high threshold like 95% and stable period before we ever claim success. This is like an “optional stopping guardrail” – if we haven’t hit that threshold, we keep observing (like not ending an experiment until enough evidence of balance).

8.6.7 Missing data, gaps, and dropouts

Missing data is unavoidable, but we handle it in a model-based way to avoid bias. We do no simple imputation on raw signals. Instead:

- Our state-space model naturally treats missing observations by not updating with that channel at that time – essentially the prediction for F carries on with increased uncertainty. This is better than filling with zeros or last value, which could bias things. In our Kalman filter, if a channel is missing at time t , we just skip its update; F_t is then predicted from F_{t-1} with larger variance.
- If an entire channel drops out permanently (like a sensor fails from day X onwards), then our uncertainty post-day X increases because one less source informs F . We quantify that by the posterior variance – it grows. We flag intervals of high uncertainty explicitly: e.g. if posterior SD of F_t exceeds some threshold (like >0.7 HCU for >6 hours), we gray out that region or indicate it as “no confident inference here.” This means if someone forgot to wear their device for a day, the ledger band will widen a lot and we might tell them “we can’t be sure about this day’s balance.”
- For structured missingness (like we know they don’t wear device during sleep or remove for shower at 7am daily), we incorporate that knowledge as covariates or known intervals so that the model doesn’t treat it as random. E.g., we might model an “off-device” period by allowing the process noise to differ or simply by acknowledging those gaps so we don’t overinterpret them. This prevents attributing significance to every gap – some missingness is just routine (not a breakdown in measurement quality).

The key is, missing data widens credible intervals – we let it. If someone asks “what happened on Tuesday when data was missing?”, answer is “we’re uncertain – could be anywhere in this broad range,” and the ledger reflects that.

8.6.8 Visualizing uncertainty

We present uncertain trajectories in ways that make reliability intuitive without overstating precision. Every summary of F_t or the cumulative ledger is accompanied by uncertainty bands and explicit annotation of calibration events.

- Uncertainty bands: The latent F_t over time is summarized as a median estimate with corresponding uncertainty intervals (e.g., 50%, 80%, 95% credible intervals). The same approach applies to the cumulative ledger. These intervals make it clear when uncertainty widens (for example, during data gaps) and when it tightens (after dense measurement or anchor calibration).
- Anchor annotations: When an anchor is applied (e.g., pain, chills, exclusion), the timing is explicitly recorded. If recalibration occurs (e.g., device linking or anchor update), that event is documented in the timeline. Periods where data are not fully comparable (such as immediately after a device change that has not yet been linked) are explicitly annotated in the record. This ensures that shifts in estimates are not mistaken for shifts in lived experience.
- Counterweight accounting: When F moves above baseline (positive contribution) or below baseline (negative contribution), the cumulative contributions are tracked separately and reported alongside their uncertainty. This makes clear not only the direction of drift but how confidently that drift is estimated.
- Cross-context comparisons: When comparing measurements across devices or settings (e.g., lab vs. home), agreement is evaluated with uncertainty intervals. If intervals overlap substantially after linking adjustments, alignment is considered adequate; if not, the discrepancy is reported rather than hidden.
- Traceability: All summaries specify the HCI/HCU version, device details, anchor dates, and model version used to generate the estimates. This enforces traceability and allows any anomaly to be audited against the exact measurement and modeling configuration.

The guiding principle is simple: no ledger value is presented without its uncertainty. Balance, neutrality, or divergence are always reported as probabilistic statements grounded in documented calibration and model assumptions.

8.6.9 Sensitivity and stress tests

We also test how robust the ledger is to various hypothetical perturbations:

- Channel ablation: We recompute the entire ledger leaving out each channel one at a time. Does the final neutrality probability or total HCU change drastically if, say, we ignore self-report or drop HRV? If dropping one channel changes conclusion, then our instrument is too dependent on that channel and results must be caveated. Typically publish an appendix figure: "Remove each channel: effect on final p_neutral and on main outcome metrics." If none of the removals flips the sign or moves p_neutral beyond a threshold, we feel more secure.

- Specification curve (multiverse analysis): Try various reasonable preprocessing or modeling choices: e.g. different ways to detrend signals, different priors on process noise, including vs. excluding minor covariates. For each, compute the final ledger median. We then show the distribution of those outcomes (say median ledger ranged from -0.5 to +0.2 across 100 specs). If that IQR is narrow and straddles similar values, good. If some choices lead to significantly different conclusions, we must report that or stick to the more conservative approach.
- Rival models: Compare our LoF/QS integrated model to alternative models (like an RL model or free-energy principle model that might also predict behavior). We check if our model's ledger uncertainty is narrower or predictions better. If a rival model explains the data equally but yields a different ledger path, we have to acknowledge model uncertainty (maybe the “true” ledger is model-dependent). If ledger results are robust across model types, more confidence.
- Heavy-tailed stress: We might swap out normal error assumptions with a heavy-tailed distribution (like Student-t with df=4) to see if outlier points were unduly influencing the ledger. If neutrality conclusion holds even with fat-tailed noise (which downweights outliers less), then it's not an artifact of assuming Gaussian.

These tests help us ensure that our pleasant story of ledger balancing isn't an artifact of a particular analysis decision. If any test shows fragility, we either refine the instrument (perhaps add a channel to stabilize) or at least flag the uncertainty (e.g. "if physiological signals are ignored, the ledger would not appear neutral").

8.6.10 When uncertainty forces humility (fail patterns)

Finally, we delineate cases where, despite all efforts, the uncertainty is too high to draw strong conclusions about fairness or balance. In such scenarios, we commit to downgrading our claims and we will *not* declare a ledger balanced if:

- The 95% CI of the final ledger straddles both positive and negative and the probability of neutrality is below, say, 0.8. E.g., if CI is [-2, +1] HCU and p_neutral = 0.5, it would be irresponsible to call that balanced or not – it's inconclusive. We'd explicitly state that.
- If more than 40% of the ledger area since last anchor is under high-uncertainty bands (meaning for large chunks we didn't really know the sign of F), we refrain from strong claims. Essentially, “the data is too noisy for too long to be confident.” We might call for more data collection.
- If doing the sensitivity tests yields different signs – e.g. in some channel-omitted scenarios the ledger is positive, in others negative – then any conclusion about fairness could flip with unseen factors. We then either present results as

contingent (“if one trusts physiology, result is X; if not, result is Y”) or we hold off on conclusions until improved methods come.

- If cross-site draws (when comparing multiple persons or groups in a fairness calculation) form a bimodal distribution *linked to a methodological factor* (like all draws where device A was used give higher outcome than device B, creating two clusters), that indicates an unresolved measurement difference. We wouldn’t confidently pool those; we’d treat them separately.

In all such cases, the answer is to report prominently the uncertainty and avoid overstepping. For example, rather than saying “The law of fairness held for this patient,” we’d say “Data were inconclusive to determine if compensation was achieved; additional monitoring or improved measurement is needed.” We might propose remedies: maybe re-anchor, fix device issues, or collect more data in critical periods.

Takeaway: The ledger is not a single line on a chart – it’s a distribution of possible trajectories. By embracing that, we present conclusions in terms of probabilities and confidence, which is scientifically honest. When the evidence warrants, we can say a life likely balanced out (with X% confidence). When evidence is weak, we refrain from claiming victory or failure of LoF – we explicitly call it a data limitation. This way, neutrality (or lack thereof) becomes a *quantified* statement – “there’s a 97% chance of neutrality” is a strong endorsement, whereas “there’s a 60% chance” means we need to observe longer or improve measurement before judging.

No ledger claim goes unqualified by its uncertainty. This protects from false negatives/positives in testing the LoF. If LoF is true but our instruments are noisy, we won’t mistakenly reject it – we’ll say “can’t tell yet.” If LoF is false in a subtle way, we also won’t erroneously confirm it without tight error bars. This rigorous approach makes any final declaration – say, at someone’s end-of-life ledger – something that others can trust because it comes with transparent margins of error and stated confidence.

8.6.11 Where we go next:

With uncertainty now explicit and carried into the ledger, we leave measurement and turn to identity. Part V asks whose ledger we are closing: how to count one stream, two streams, or a paused stream—and how those calls affect any claim about fairness.

Part V — Identity and Edge Cases

A soldier wakes up from a months-long coma with no memory of the injury that put him there. Across the world, a person with dissociative identity disorder lives with alternating personalities that each hold different memories and feelings. In a neuroscience lab, a split-brain patient (with the connection between brain hemispheres severed) performs tasks that make it seem like two separate minds inhabit one body. What counts as “one life” when it comes to fairness? If the Law of Fairness guarantees each unified conscious stream its own neutral ledger of experience, we need to be crystal clear about what that *stream* is – especially when identity isn’t straightforward. Notation: $L(t)$ denotes the latent life ledger (the time-integral of momentary felt valence $F(t)$); $\hat{L}(t)$ denotes its estimate from HCI/HCU, reported with uncertainty. Part V tackles these questions of identity and continuity. We define the unit of accounting not as a body or a soul, but as a *stream of conscious access*. This lets LoF apply even when the usual markers of identity get fuzzy. In plain terms: a “life” under LoF corresponds to a continuous, unified conscious perspective. But what if that perspective splits or pauses? Here we explore how the fairness ledger handles edge cases like splits, merges, and gaps in consciousness.

First, we confront the idea that one physical person could host multiple conscious streams or, conversely, that one stream might span what we usually think of as two people. While such cases are rare, they are incredibly important for a theory that claims no exceptions. LoF is meant to hold for every lawful conscious life. So we ask: *When a mind splits, does each part get its own ledger? When a mind merges or a lapse ends, do ledgers fuse back together?* The answer we develop is grounded in the concept of conscious access. If two streams of experience truly have separate awareness (unable to directly share memories or experiences), then LoF treats them as separate ledgers – separate “lives” for fairness purposes – even if they happen to reside in one skull. For example, in a classic split-brain scenario, if each hemisphere has its own independent experiences, each would need its own fair ledger (at least until unified experience is restored). In dissociative identity disorder, if different alters experience life disjointedly (with barriers between their memories), LoF would conceptually assign each alter their own stream-of-experience ledger while they are functionally separate. The moment the streams unify – say, through therapy or when one alter is dominant – the ledger accounting would merge or switch accordingly. Dominance alone is not a unity call; merge/switch rules trigger only when preregistered access criteria indicate restored bidirectional sharing. Similarly, during a coma or deep anesthesia, a person’s conscious stream is paused (no experiences to tally), and it resumes when they regain

consciousness, picking up the same ledger rather than starting a new one. Operationally, we classify coma/anesthesia as a pause only when the Unity/Access criteria indicate no unified experience; if access intermittently returns (dreaming, covert awareness, MCS), the ledger accrues during those windows with explicit uncertainty rather than being assumed zero.

All of this might sound philosophical, but it has practical implications for testing the law. It means our experiments and data must pay attention to unity of consciousness. If a person has fragmented consciousness, we might actually be observing multiple ledgers in play. Part V lays down how we handle these issues so that we keep our fairness claims testable and fair. There's no need for mystical notions of an eternal "self" – we use observable criteria (like communication between mental states, memory continuity, and integration of information) to decide what constitutes one stream. This approach lets us include edge cases *without* breaking the rules or resorting to special exceptions. It also ensures that when we say "everyone ends up with a neutral ledger," we're careful about defining "everyone." It's not every body or every legal person, but every unified conscious trajectory.

Finally, this part is also about making sure our metrics are universally meaningful across those different trajectories. Even once we've decided whose ledger is whose, we face the question: does our measure of comfort (HCl and related indices) mean the same thing for all streams of experience? If not, we'd be comparing apples to oranges. For comparability methods (measurement invariance), see Chapter 8 — 'Same Scale' Across People and Places. Part V focuses on identity and ledger continuity: defining one stream vs. two, handling pauses, and specifying split/merge rules. Part V, in summary, shores up the definition of "one life, one ledger" and makes sure the yardsticks we use are truly universal.

What this Part will do for you:

- One ledger per conscious stream: A clear explanation of how LoF defines an individual "life" or unit of fairness. We'll see why we use *streams of conscious access* rather than an essentialist notion of self – allowing the theory to cover cases like split brains, multiple personalities, or long unconscious gaps without breaking.
- Handling splits, merges, and gaps: Insight into how the system handles edge cases of identity. We'll walk through thought experiments (and real examples) of consciousness splitting or pausing, and show how fairness ledgers would partition or reconnect accordingly. This ensures the Law of Fairness can be

applied consistently even when someone's identity is ambiguous or changes over time.

- Ledger rules for everyone: Practical criteria for assigning one ledger per unified stream, preserving ledger continuity through pauses, and applying pre-registered split/merge rules in edge cases.
- Adjudication under uncertainty: When unity is ambiguous, apply dual protection and local-only claims until follow-up clarifies the call, with no retroactive combining after outcomes are known.

Chapter in this Part:

- **Chapter 9 — Unity of the Stream** - Sets the operational criteria for “one stream vs. two,” explains how pauses (sleep, anesthesia, some disorders of consciousness) are handled as the same ledger, and addresses difficult cases (split-brain, DID, AI, organoids) with blinded adjudication and preregistered thresholds. The goal is practical guidance: enough structure that different teams can make the same call from the same packet of evidence.

Where we go next:

Part V opens by asking a precise question—when do multiple processes add up to one experienter? Chapter 9 starts with that call and the minimal evidence needed to make it consistently and humanely.

Chapter 9 — Unity of the Stream

We define the Unity of the Stream—when multiple processes count as one experienter for fairness accounting. We introduce a reader-facing Unity Index (Plain Speech) to decide ‘one vs. two’ from observable signs, then show how pauses, splits, and merges affect ledger continuity and the Queue System (QS). Comparability across people is treated elsewhere (see Chapter 8).

Let’s put it concretely. Imagine part of HCI is a self-reported mood score. Person A might be a tough grader (they say they’re “5/10” even when they’re fairly comfortable), whereas Person B might be more lenient (they’ll say “8/10” under similar comfort). If we took those self-ratings at face value, HCI could end up underestimating A’s comfort relative to B’s. Measurement invariance testing helps detect this kind of issue. We ask: *Is the HCI construct interpreted in the same way by different individuals or groups?* Technically, we test whether the underlying factor structure of HCI is consistent across groups (that’s called configural invariance), whether the contributions of each component to HCI are equal across groups (metric invariance), and whether even the absolute levels are comparable (scalar invariance). You can think of it like calibrating thermometers: two thermometers should read the same temperature in the same room. Configural invariance checks that both thermometers have the same basic shape of measurement (they respond to temperature changes in a similar pattern). Metric invariance checks that a 1-degree change moves both needles equally (same units). Scalar invariance checks that they both read exactly the same at a known reference point (no offset or bias). If any of these checks fail, then comparing readings (or HCI scores) directly could be misleading.

In practice, we apply multi-group statistical tests to HCI data to probe these questions. For example, we might split our data by demographic group or by site of care and see if a single HCI model fits all groups well. Configural invariance means all groups appear to have the same basic HCI structure – they weigh the various inputs (pain, mood, etc.) in qualitatively similar ways. If this holds, it suggests everyone conceptualizes “comfort” similarly in terms of the components we’re measuring, which is a great start. Next comes metric invariance: we statistically constrain those component weights (also called factor loadings) to be equal across groups and then check the model fit again. If the fit now worsens significantly, it indicates at least one component of HCI means something different across groups (for instance, perhaps pain score is *more* impactful on overall comfort for one group than another). In our analysis, we did find some wrinkles – HCI did not achieve perfect metric invariance across all groups. This isn’t unusual in real data (people are diverse!), but it carries an important implication: we cannot directly compare

absolute HCl scores between certain groups with full confidence. In other words, a score difference might partly reflect measurement bias, not a true comfort difference.

What do we do with that knowledge? LoF is about fairness, so we err on the side of caution. If strict invariance fails, we don't blindly compare, say, an HCl of 70 for group X to a 70 for group Y and assume they represent the same well-being. Instead, we restrict some of our fairness analyses to within-person comparisons or use carefully calibrated adjustments for cross-group comparisons. "Within-person" means we focus on how an individual's HCl changes relative to their own baseline. That approach is usually safe even if different people use the scale differently, because each person serves as their own reference point. For example, if Person A's HCl goes from 50 to 60, we know *that individual* improved, regardless of how A's scale might compare to B's. The Law of Fairness can accommodate these findings by focusing on ensuring each person's improvement or maintenance of comfort over time meets the targets, rather than forcing direct one-to-one score equality between different people. Additionally, if needed, we adjust our measurement model – sometimes using a technique called partial invariance, where we allow certain components of HCl to vary by group if the data strongly suggest it – or we implement group-specific calibration offsets. But the simplest solution (and the one we adopt under LoF whenever appropriate) is to not overclaim across-group comparisons when metric invariance doesn't hold. In practice, this means LoF guarantees are often framed in a *within-person* sense: everyone should experience the relief they need relative to their own starting point, and no one should be left in avoidable suffering in their own context. We are careful when saying someone is "ahead" or "behind" another on the HCl scale unless we've accounted for any measurement bias that could be in play.

Despite these hiccups, there's a silver lining. The fact that we usually achieve configural invariance across many groups suggests HCl is tapping into the right core concept of comfort for everyone. The differences that caused metric invariance to falter might be small or limited to certain sub-components. We will report exactly where our measure misbehaved – for example, maybe the "anxiety" component of HCl carried slightly different weight for older adults versus younger adults. Knowing this, we can either refine HCl (say, age-tailor that component) or simply be mindful of it in our analyses. The key point is that LoF's fairness guarantee doesn't crumble because of these measurement quirks; it just means we apply the fairness rules with a bit of nuance. We might guarantee that *each person* will have their comfort raised to a high level relative to their own baseline (a meaningful form of fairness in itself), rather than guarantee that a raw HCl number is identical for all individuals. In essence, we won't let a measurement bias

masquerade as a real unfairness – we either correct the bias or adjust our interpretation so that we continue to only make true comparisons.

What you'll get from this Chapter:

- The importance of invariance: A down-to-earth explanation of what measurement invariance is and why it matters for fairness. If we're going to say "everyone is at HCl 80 at end-of-life," we need to ensure an 80 means the same comfort level for all. You'll see examples of how interpretations can differ between people and why that could lead to false conclusions if not checked.
- How we test for invariance: An overview of the testing process – moving step by step from configural to metric to (if possible) scalar invariance. We'll describe how multi-group confirmatory factor analysis works in concept (without heavy math): basically, fitting our HCl model separately to different groups, then fitting it again with cross-group constraints, and seeing if the fit worsens significantly when we force a "one-size-fits-all" assumption. We'll mention common criteria in passing (like chi-square difference tests or changes in fit indices like CFI) but keep the discussion non-technical in this intro.
- What our tests found: The results of our invariance tests for HCl. For instance, you'll learn that configural invariance held – the overall factor structure was consistent across groups – a good sign. However, metric invariance partially failed: some factor loadings could not be equal without model fit deteriorating. We explain in simple terms what that means (e.g., "certain sub-scales of HCl carry slightly different weight for different groups, so the 'unit' of HCl isn't identical across everyone"). If we tested scalar invariance, we'll report on that too (though if metric didn't hold, scalar invariance likely didn't either).
- Implications for fairness analysis: Most importantly, we translate those invariance results into practical policy. You'll see how LoF's enforcement rules adjust when full invariance isn't met. This might mean focusing on within-person fairness – ensuring each individual's comfort improves to meet criteria relative to their own history – rather than insisting on direct cross-person comparisons that could be skewed by measurement differences. We emphasize that this adjustment still honors LoF's spirit: no one is left in avoidable suffering. If measurement quirks make direct comparisons tricky, we simply change how we evaluate fairness (for example, guaranteeing that *each person* reaches a high comfort level, even if their "80" isn't exactly someone else's "80").
- Transparency and future calibration: We conclude by noting that discovering non-invariance is actually a success of our rigorous approach. It highlights where we can improve our metrics or apply calibration. We might mention that further

research could allow us to tweak HCI to achieve stronger invariance (for instance, by adding group-specific adjustments or using anchoring vignettes to re-calibrate subjective scores). For now, we proceed with eyes open: our fairness decisions are made with an awareness of these measurement limits, and we communicate them clearly. In short, we don't let a measurement bias hide beneath the rug—we either fix the bias or work around it, ensuring that what we call "fair" is truly fair.

Subsections in this Chapter:

- **9.1 Conscious Access: One Stream or Two** - Sets the operational criteria for deciding whether experiences belong to a single conscious stream or to multiple streams, focusing on global access, integration, and coherent control. Establishes why this call is the prerequisite for assigning a single ledger.
- **9.2 The Unity Index (Plain Speech)** - Introduces a pragmatic scoring tool—built from observable signs like cross-talk, shared memory, coordinated action—to adjudicate unity without jargon. Explains inputs, thresholds, and how the score is used consistently across cases.
- **9.3 Pauses: Sleep, Anesthesia, Coma** - Clarifies when the ledger accrues and when it pauses. Dreamless sleep, deep anesthesia, and true coma suspend accrual; REM/dreamful periods may contribute entries. Details bedside signals and device markers used to detect pauses and brief returns.
- **9.4 Split-Brain and DID** - Handles fragmentation and reunion. When streams run in parallel (e.g., split-brain) or alternate with amnesia (some DID presentations), ledgers branch; when unity returns, ledgers fuse by summation with uncertainty carried forward. Fixes rules up front to avoid post hoc edits.
- **9.5 AI and Brain Organoids** - Extends criteria cautiously beyond humans. Only entities meeting predeclared evidence for unified consciousness (access, integration, control, valence) get a ledger; otherwise they are excluded. Outlines safeguards and conservative inclusion standards.
- **9.6 Research Notes: Blinded Adjudication and Thresholds** - Specifies blinded procedures, inter-rater checks, and pre-registered thresholds for unity/pauses/splits. Details how ambiguity is handled, how disagreement is resolved, and how decisions are logged for audit.

Where we go next:

We now move from the overview to the test itself. 9.1 lays out the operational access criteria—what we must see (and not see) to classify one stream versus two—and why that call is the prerequisite for assigning a single ledger.

9.1 Conscious Access: One Stream or Two

A single brain can host one conscious stream or more than one—but not every case of parallel processing counts as two minds. This section turns “one vs. two” into an experimental question by focusing on conscious access: do information states in subsystem A become available to subsystem B (and vice versa) for report, valuation, memory, and action within a short temporal window? If yes, they belong to the same stream. This resonates with global workspace theory (Baars, 1988): conscious content becomes available only when broadcast widely across the brain, whereas information that fails to ignite such broadcasting remains isolated (unconscious). According to Integrated Information Theory (Tononi, 2008), one can quantify unity of consciousness by a parameter Φ that measures information integration: a truly unified stream should exhibit high Φ , whereas two disconnected streams would have much lower combined Φ . To avoid symbol drift, we denote IIT’s integration metric as Φ_{IIT} ; Φ in LoF is reserved for feasibility-of-compensation in QS analyses. In practice, measuring Φ_{IIT} could help test whether a system is operating as a single conscious unit or as two separate ones (Tononi, 2008). In this book, these theory links are motivational only: operational unity is decided by the access/perturbation evidence below, and Φ_{IIT} is treated as a non-decisive proxy (not a required computation) in real brains. If not—and if the unavailability is sustained and bidirectional—we treat them as separate streams with separate ledgers and separate QS menus.

9.1.1 The access test (operational core)

We define conscious access as present-to-near-future availability of content across the control architecture. Availability is demonstrated when briefly presented or internally generated content:

1. Alters action in another module (e.g., a somatic cue changes a visual choice).
2. Enters report (verbal, motor, or BCI) outside its origin module.
3. Is re-valued by systems that did not generate it (vmPFC/OFC update).
4. Is remembered by modules that did not encode it initially.

The time window is short (hundreds of milliseconds to a few seconds) because LoF and the QS adjudicate options on behaviorally relevant horizons. Mere offline consolidation hours later does not establish same-stream status.

Decision rule: If two putative parts show reliable, bidirectional access on ≥ 2 of the four criteria above (with preregistered thresholds), we classify them as one stream. If they fail

all criteria across multiple paradigms, we classify them as two streams. Mixed results trigger ambiguous status and dual protection.

9.1.2 Parallel processing is not a second stream

Humans can dual-task, mind-wander during driving, or solve a problem while brewing coffee. These are parallel processes, but they typically share access: a sudden hazard on the road interrupts the inner monologue; a sharp somatic pang shifts visual search. We therefore distinguish:

- Parallel-but-accessible: Content in process A can intrude on process B quickly; a global workspace ignites and broadcasts. This remains one stream.
- Partitioned-and-inaccessible: Content in A cannot intrude on B and vice versa, even when stakes are high and signals are strong; perturbation does not restore access. This suggests two streams.

9.1.3 Four families of access paradigms

We use at least two families per evaluation to avoid task idiosyncrasies.

A. Cross-channel conflict and override

- Interoceptive–exteroceptive conflict: Present a strong interoceptive cue (e.g., inspiratory load) while a visual decision is underway. In one stream, conflict slows and biases the visual choice and is later reported as a single integrated difficulty. In two streams, the visual choice proceeds unmodulated while the interoceptive module shows strain without cross-effect.
- Action override: A covert motor plan detected by EMG should be cancellable by a late-arriving value update if access holds (rlFG/ACC signatures). Failure to cancel across the partition suggests separate streams.

B. Masked primes and global broadcasting

- Subliminal cue in A, report in B: A masked word presented to the right visual field (left hemisphere) should primarily bias the right hand; cross-side bias or recognition by the other hand is the access signature. If hemisphere access is severed, any cross-side bias should vanish and remain unreportable across sides.
- Ignition tests: Brief, near-threshold stimuli that typically trigger fronto-parietal ignition should ignite globally in one stream; in two streams, ignition remains local.

C. Memory carryover and policy continuity

- Counterfactual learning: Train a preference or “if–then” policy in module A; test transfer in module B minutes later. One stream transfers; two streams do not, absent re-teaching.
- Episodic tagging: Tag an event with a subtle cue (odor, tone) in A; test whether B’s later recall shows the same tag effect.

D. Perturb-and-measure connectivity

- TMS–EEG effective connectivity: Perturb a node in A (e.g., left premotor) and measure evoked responses in B (e.g., right parietal). One stream shows reliable causal spread; two streams show attenuated or absent propagation despite matched stimulation.
- Endogenous perturbations: Induce brief arousal bursts (startle, breath-hold) in A; B should show time-locked signatures if access holds.

Thresholds (preregistered): e.g., cross-module reaction-time interference ≥ 20 ms with $p < .01$ and Bayesian $BF > 10$; TMS-evoked potential in B at ≥ 3 SD above baseline on $\geq 60\%$ of trials; memory transfer $d \geq 0.3$. Here d is a placeholder for the preregistered standardized transfer metric (e.g., d' in signal-detection tasks); the exact choice is fixed per paradigm. These are example magnitudes; exact SESOI, α levels, and Bayesian criteria must be fixed in preregistration and justified by pilot data or meta-analytic priors for each cohort.

9.1.4 Applying the test to canonical cases

Split-brain (complete callosotomy). These predictions are conditional on the completeness of disconnection and residual pathways; the preregistered access tests—not the diagnostic label—determine the unity call. Prediction under one stream with subcortical/collicular access: cross-field masked primes still bias contralateral actions; startle/arousal propagates; some delayed transfer of counterfactual policies occurs. Prediction under one stream with sufficient residual interhemispheric or subcortical access: cross-field masked primes bias contralateral actions above preregistered SESOI; startle/arousal shows time-locked cross-hemispheric signatures; delayed transfer of counterfactual policies exceeds chance under blinded testing. Prediction under two streams: cross-field biases vanish; TMS–EEG shows minimal interhemispheric spread; each hand can learn a policy the other cannot access even with incentives. QS should partition menus: the left-hand stream’s admissible set can tilt toward repair independently of the right-hand stream’s.

Dissociative identity disorder (DID). One stream if amnesic barriers collapse under incidental probes (e.g., odor tags, priming) and counterfactual policies trained in one state transfer when incentives align. Two streams if barriers persist across probes and perturbations fail to cause cross-access. Clinical implication: local ledgers during persistent partition; QS acts per alter. Recovery prediction: as therapy restores access, QS effects become more global and counterweights appear across formerly isolated states. Dual-tasking and mind-wandering. One stream: robust global interrupts; masked primes hop across tasks; memory tags bind. Two streams would require sustained inability of hazard signals to intrude, which would itself be pathological.

9.1.5 Ties to the Queue System (QS)

QS attaches to the unit that can compensate itself, i.e., the unit with access sufficient to support policy-level counterweights. Consequences:

- In one stream, admissible options are computed against the whole ledger $L(t)$. A late-life reconciliation call may feel compulsory because it optimizes global compensability.
- In partial partitions, QS may weight options differently across submodules (e.g., left vs. right hand) but still coordinate counterweights with delays.
- In two streams, QS partitions. A harmful option in stream A may be pruned even if stream B could hypothetically compensate, because no access means no guarantee. This yields testable dissociations: high- Φ repair options rise in one part while neutral or indulgent options in the other lose “stickiness,” even when utilities match.

9.1.6 Practical protocol (clinics and labs)

1. Screen and consent. Explain that classification affects how we analyze fairness and design care; adopt conservative defaults.
2. Choose two families of paradigms suited to the case (e.g., masked primes + TMS–EEG for split-brain; memory carryover + access tasks for DID).
3. Preregister thresholds and nuisance regressors (arousal, effort, medication).
4. Run perturb-and-measure first, then access tasks, then memory/policy transfer.
5. Classify: one, two, or ambiguous. For ambiguous, apply dual protection and local-only claims until follow-up clarifies.
6. Document Unity Index components for transparency and future pooling.

9.1.7 Fail conditions for this section

- Cross-compensation without access: Robust evidence that counterweights in B regularly repair harms in A despite failed access tests and absent causal spread would challenge QS-unity coupling.
- Access without broadcast: If access metrics pass but global ignition and policy continuity repeatedly fail, we refine thresholds or require triangulation with additional domains.

9.1.8 Everyday feel (why readers will recognize this)

People notice unity when a sudden bodily jolt interrupts a thought, when a memory recolors a choice seconds later, when an outside sound grabs attention and changes a plan. They notice fragmentation during extreme fatigue, panic, or dissociation, when parts of the self feel sealed off. The experiments above formalize that intuition: intrusion, influence, and integration mark one stream; sealed channels mark two.

Takeaway. Conscious access is the passport stamp for stream membership. With perturbation, access tasks, and memory transfer, “one vs. two” becomes measurable. That decision fixes whose ledger LoF balances and where QS must do its work. Next, we compress these ingredients into a reader-facing Unity Index (9.2) and then its technical form (9.3).

9.1.9 Where we go next:

With access defined and tested, we turn the criteria into a practical score. 9.2 presents the Unity Index in plain speech—a weighted, preregistered combination of observable signs—that lets independent teams make the same unity call and document how they made it.

9.2 The Unity Index (Plain Speech)

The Unity Index is a pragmatic score that helps decide “one stream or two” from observable signs. It is not a metaphysical claim and it is not a black box. It is a preregistered checklist, scored under blinding where feasible, that aggregates evidence about cross-talk, shared memory, coordinated control, and self-model coherence. The aim is simple: given the same packet, different teams should make the same call.

9.2.1 What the index is (in everyday terms)

The Unity Index is a weight-of-evidence score. Each domain contributes yes/no or graded indicators that a single subject is accessing, integrating, and controlling the same pool of information. High scores mean “one stream” is strongly supported; low scores mean “two streams” (or “no stream”) is more likely; a middle band is classified as ambiguous with protective defaults.

9.2.2 Observable signs we look for

Cross-talk between subsystems: information injected in one channel reliably influences processing in another (e.g., a tactile cue shaping visual search). Shared memory across domains: content experienced “here” later appears “there” without special training or external prompts. Coordinated control: actions reflect integrated goals (stopping one hand because the other just learned a hazard). Coherent self-model: first-person reports (when available) and behavior point to a single “I” that owns perceptions and actions.

9.2.3 How we score (transparent and preregistered)

Each domain is scored with simple rubrics (e.g., present/absent; weak/moderate/strong) tied to specific tasks and thresholds written in advance. Weights reflect evidential strength and feasibility (e.g., cross-modal transfer > self-report phrasing). The composite is a weighted sum or tally that maps to three decisions: One, Two, or Ambiguous. All thresholds and tie-breaks are preregistered.

9.2.4 The packet: what scorers actually see

To keep decisions replicable, scorers receive a fixed packet: task descriptions, de-identified traces (behavioral timings, simple physiology, stimulus logs), minimal narrative context, and any first-person reports collected under standard prompts. They do not see hypotheses, clinic notes, or identities unless consent and design require it.

9.2.5 Thresholds and conservative defaults

Classify as “One” only when pre-specified criteria across multiple domains are met. Classify as “Two” when contrary evidence is strong and coherent. Otherwise classify

“Ambiguous,” use protective defaults (separate ledgers for high-stakes decisions), and plan follow-up. Relief is a systems variable; comfort and dignity override data collection.

9.2.6 Keeping it honest (blinding and drift control)

Where possible, scoring is blinded to diagnosis and hypothesis. Raters train on shared examples and re-calibrate periodically. We log disagreements, compute inter-rater reliability, and update rubrics only between studies. Any deviation from the preregistered thresholds is recorded in a change log.

9.2.7 Fail patterns worth watching

If independent teams cannot agree on the same packet, the index is not yet decision-grade. If domains regularly point in opposite directions (e.g., cross-talk present, but no shared memory under clean designs), the construct is unstable. If thresholds can be tuned post hoc to produce any desired answer, the tool fails.

9.2.8 From index to ledger rules

Unity calls drive accounting. ‘One stream’ means one ledger through pauses; ‘Two streams’ means separate ledgers with clear, pre-written rules for splits and merges; ‘Ambiguous’ keeps ledgers separate until clarity (see 9.1.6 for ambiguous-case safeguards). No retroactive combining after outcomes are known.

9.2.9 Where we go next:

The plain-speech index now in hand, 9.3 applies it to pauses—sleep, anesthesia, coma—showing how we preserve ledger continuity when access collapses and returns, and what would count against that continuity.

9.3 Pauses: Sleep, Anesthesia, Coma

Not all breaks in responsiveness are breaks in the stream. In this section, we distinguish a pause (a reversible suppression of consciousness with later re-binding of access) from an end (the permanent loss of the stream, i.e., death of mind). We show how to treat sleep, anesthesia, and disorders of consciousness (DoC) within the LoF framework and how the Queue System adapts during these states.

9.3.1 The guiding idea: pause vs. end

- Pause (suppression): Conscious access is downregulated, sometimes all the way to zero outward responsiveness, but diachronic identity is *preserved*. When the system returns (wakes up, regains consciousness), it behaves as the same chooser it was before – its policies, memories, and valuation patterns re-bind to the prior self. The life ledger continues across the gap in experience.
- End (termination): Access does not return and cannot return given the physical state of the system. There is no capacity for future counterweights or further unified experience. The ledger closes at the moment of ‘death of mind’ (see 3.5 ‘The Death of Mind’ for the formal definition).

We operationalize pause vs. end by tracking the Unity Index and corroborating structural or physiological criteria over time (9.2), rather than relying on any single proxy. A pause shows loss of consciousness (low Unity Index, impairment in two or more index domains) followed by later restoration of those domains upon recovery.

9.3.2 Sleep: a planned, rhythmic pause

Sleep is nature’s safest daily suspension of access. Here “suspension of access” is stage-dependent: truly dreamless deep sleep is modeled as $F(t)\approx 0$, while dreamful periods still contribute experience entries even though the body is offline. LoF treats sleep as ledger-neutral downtime with two special roles:

- Maintenance and rebalance: Sleep provides physiological and mnemonic “housekeeping” that reduces future hedonic load – essentially a repair gain. (For example, deep sleep helps reset stress systems and heal tissue, preventing negative ledger accrual the next day.)
- Counterweights via dreams: REM sleep dreams can simulate threatening or emotionally significant scenarios at low bodily cost. This serves as a kind of off-line practice or emotional release, providing low-cost corrections when waking life has accumulated imbalance (see 5.4 on dreams as “the night workshop”).

Unity and access across sleep stages:

- NREM (stage N2/N3): Greatly reduced global broadcasting; sensory inputs are largely gated out; the brain shows slow-wave activity. In terms of the Unity Index, a person in deep NREM scores very low (access is largely lost), but crucially diachronic identity is preserved – when they wake, they re-bind to the same self and ledger.
- REM: Dream-rich state with partial restoration of affective processing; treat as low-cost counterweight opportunity (see 5.4), with Unity Index increasing relative to deep NREM while ledger continuity remains intact.

Predictions for LoF/QS during sleep:

- Horizon effect: On nights following days of high negative ledger drift (lots of unmet needs or pain), the pressure to enter sleep increases and REM density (the intensity or frequency of REM episodes) goes up. In other words, QS would tilt the system toward more reparative sleep options – making you sleep more deeply or dream more intensely to catch up on repair.
- Dream content tilt: After very stressful or imbalance-heavy days, dreams should contain more mastery, reconciliation, or exposure themes (e.g. overcoming challenges or making amends), whereas after long periods of monotony or self-denial, dreams may skew more indulgent or wish-fulfilling. These are testable covariation claims about theme frequencies under blinded coding, not assertions of fixed symbolic meanings.

Measurements:

- EEG signatures: Slow-wave activity (SWA) in deep NREM sleep correlates with prior wake effort and negative load (the more depleted or stressed you were, the more SWA you get). REM sleep *density* (frequency/intensity of rapid-eye-movement periods) tracks the demand for counterweight dreaming.
- Morning re-binding: After a normal night's sleep, a person's Unity Index should rebound to baseline – their policy continuity and interoceptive attunement return to pre-sleep levels. (If someone wakes up still fragmented, something is wrong – e.g. sedative drugs or a neurological issue.)

Notably, if our theory is right, certain failures should not happen: if someone accumulates a large negative ledger (e.g. extreme stress), it should affect their sleep architecture. If chronic high stress or negative drift *does not* alter sleep pressure or architecture (after controlling for other factors), then our assumed QS-sleep coupling is

weakened. Similarly, if detailed content analysis shows no systematic shift in dream themes based on ledger history, then the notion that dreams serve as “counterweights” is undermined.

9.3.3 Anesthesia: engineered pause with dose–response control

General anesthesia is a pharmacologically induced suspension of access. LoF treats adequate anesthesia (sufficient depth to ensure unconsciousness) as a pause *if and only if* two conditions are met:

1. Unity collapses during anesthesia: The patient shows no evidence of conscious access — no “talk-through,” no global broadcast of information, no coherent policy-directed behavior. (In practice, this means unresponsiveness, no report, and suppressed brain integration.)
2. Unity re-binds on emergence: When anesthesia wears off, the patient’s preferences, memories, and self-continuity reappear (allowing for a brief delirium). In other words, apart from some confusion, they pick up being the *same person* they were before the surgery.

If both conditions hold, then anesthesia did not break the stream’s identity – it just paused it.

All such markers (including PCI and EEG patterns) are treated as empirical proxies for loss/return of conscious access, interpreted alongside behavioral evidence and uncertainty.

Mechanisms and markers: Under common anesthetic agents:

- Thalamocortical gating and cortical uncoupling: Communication between the thalamus and cortex is disrupted, and frontal–parietal networks uncouple, reducing global availability of information (essentially, the “broadcast” system is shut off).
- Perturbational Complexity Index (PCI): This is a TMS–EEG-based measure of consciousness. It typically drops below the threshold for unity during adequate anesthesia and recovers upon emergence (unless the patient is experiencing the rare event of intraoperative awareness).
- EEG spectral signatures: Characteristic patterns like alpha–delta oscillations (for propofol) or burst-suppression at very deep levels serve as indicators that the brain is in an anesthetized, non-integrated state.

QS and admissible options: During proper anesthesia, QS treats the person's stream as offline for active decision-making. The person isn't making choices, so QS instead applies to the context: the perioperative environment becomes the domain of choices. In practice, that means the medical team's decisions about analgesia, temperature, noise, etc., should aim to minimize negative ledger accumulation during the pause. The idea is to avoid leaving the patient with a big pain or stress debt when they wake.

Prediction: Better control of nociception (pain signals) during surgery will result in less post-operative negative drift (the patient's ledger won't be pushed far negative by unmitigated surgical pain) and a shorter period of intense "repair" actions after surgery, controlling for surgical severity, baseline health, and medication effects. In simpler terms, if you keep the patient's body as comfortable as possible while they're under, they should have a smoother recovery with fewer compensatory needs.

Protocol for emergence checks: Before discharging a patient from the post-anesthesia care unit (PACU), one could do a quick Unity Index screen: for example, measure an auditory oddball P3 response as a proxy for global broadcast, do a quick startle-handgrip interruption test as a proxy for talk-through (does a sudden sound interrupt an ongoing action), and a policy probe (did they maintain a simple preference from pre-op to post-op). Additionally, perform a ledger continuity test: compare the patient's value-tradeoff choices (or mood/affect measures) from before surgery to after recovery. We expect the person to still be "the same chooser" once the immediate anesthetic effects (like delirium) wear off.

Fail conditions: If we repeatedly observed patients showing intact Unity (by our metrics) *during* surgical levels of anesthesia (well beyond known cases of accidental awareness), we'd have to question either our Index or the assumption about anesthetic depth. Conversely, if someone who seemingly had an uneventful anesthesia later shows long-term breaks in policy continuity (like a lasting personality or preference change), it suggests something went very wrong – possibly a hidden brain injury or that what we thought was a pause was actually an end event in terms of their original stream.

9.3.4 Disorders of consciousness (DoC): prolonged pauses, uncertain paths

"Disorders of consciousness" include coma, the vegetative state (also known as unresponsive wakefulness syndrome, UWS), and the minimally conscious state (MCS). The challenge here is deciding whether we have a prolonged pause with some hope of re-binding, or an approaching end of the stream.

Working definitions:

- Coma: Eyes closed, no sleep-wake cycles, no response to commands. (A deeply suppressed state, usually acute.)
- UWS/Vegetative State: Sleep-wake cycles return (the person may open eyes and have periods of apparent wakefulness), but there are no consistent responses to commands or signs of conscious awareness.
- MCS (Minimally Conscious State): Intermittent, reproducible signs of awareness. For example, the person might sometimes track an object with their eyes or follow a simple command inconsistently. They have *some* access, but it's very limited or unreliable.

We adapt the Unity Index for non-communicative patients using proxies:

- Talk-through proxy: Use a startle, breath-hold, or other autonomic perturbation and look for correlated cortical responses (e.g. a surge in EEG activity). If a sudden loud noise produces a measurable brain response that echoes across regions, that's a good sign of some connectivity.
- Broadcast proxy: Look for signs like an auditory oddball P3 or a mismatch negativity (MMN) in EEG – these indicate the brain is noticing and differentiating stimuli (like recognizing its own name or an unexpected tone).
- MMN can reflect preserved sensory discrimination without conscious access; by itself it is not treated as decisive evidence of awareness and is weighted conservatively relative to command-following or stronger broadcast/complexity markers.
- Use-everywhere proxy: Try mental imagery commands detectable by fMRI or EEG (the famous “tennis vs. navigation” task: ask the patient to imagine playing tennis for “yes” and walking around their house for “no,” then decode their brain activity). Also, attempt simple BCI communication if possible.
- Bound-together proxy: Present simultaneous stimuli in two modalities (e.g. a vibration and a tone together under time pressure) and look for cross-modal EEG coupling. If the brain is integrating them, it suggests some unified processing.

Classification and ethics:

- If the adapted Unity Index is mixed/ambiguous, we err on the side of assuming the person *might* be in there. We treat it as a pause with dual protection: assume there is a present or latent stream and also guard against the possibility there isn't. Practically, this means we continue full comfort measures and try to stimulate

recovery, but also avoid any irreversible harm (like we wouldn't give up on them prematurely).

- If we get clear signs of consciousness (MCS-plus or reliable BCI responses), we treat it as a pause. The ledger is considered ongoing; QS (through caregivers) should focus on actions that maximize comfort and any chance of restoring access (e.g. pain relief, meaningful auditory stimulation, presence of loved ones).
- If there's persistent absence of signs *combined* with structural findings that basically preclude return (e.g. the brainstem is destroyed, or there's diffuse cortical death), and multiple independent teams concur, then we classify it as end of mind – the stream is gone. This is essentially the clinical scenario for declaring brain death or a permanent vegetative state when truly confirmed.
- Clinically, brain death (whole-brain death) is distinct from UWS/MCS diagnoses; “end of mind” here refers to irreversible loss of unified conscious experience, which may or may not coincide with biological death.

LoF-inspired predictions in DoC trajectories:

- When a patient in DoC shows a significant re-binding event – say the first time they give a reliable BCI answer or consistently track a moving object – that should be followed by QS-driven changes in care. Specifically, we'd expect to see the care team (perhaps unconsciously) shift toward more reparative inputs: for example, once a patient starts indicating awareness, staff and family will increase things like pain management, comforting communication, playing familiar music, etc., as if some “horizon” of possible recovery has opened. (We could actually measure this as an increase in certain types of care actions logged in medical records.)
- Patients who show isolated ‘islands’ of broadcast activity (like a strong P3 response to their name) have a preregistered higher probability of later showing more complex behaviors (policy expression), controlling for injury severity and time since insult. In LoF terms, any inference about differential ledger responsiveness to positive inputs must be tested prospectively under blinded conditions rather than assumed from theory alone. In other words, even if they can't outwardly respond, their internal ledger might improve with meaningful stimuli.

Fail conditions: If we later found, through patient reports or other means, that a person had covert consciousness with zero proxy signals, that's a serious fail. For instance, suppose a DoC patient gives a detailed account after recovery and it turns out they were

having unified experiences (“streams”) for months despite us getting no MMN, no fMRI response, nothing. That would mean our proxies missed real consciousness. We’d have to add new proxies or adjust our approach because we’d been blind to an active stream. Conversely, if we often see apparent brain signals (like some EEG sign of awareness) that never correlate with any meaningful outcome (the patient never shows any further sign of consciousness or recovery), then those proxies are giving false hope and should be discounted.

9.3.5 Ledger continuity through pauses

During a pause of consciousness, the ledger state persists mathematically; accrual depends on estimated $F(t)$ over intervals where evidence supports affective contribution, with uncertainty propagated through gaps. We represent the life ledger as:

$$L(t) = \int_0^t F(\tau) d\tau,$$

where integration proceeds only over periods with defined affect estimates and carries forward uncertainty bands during gaps. The ledger state persists across pauses, but accrual depends on estimated $F(t)$. In a deep coma or under anesthesia, $F(t)$ is modeled as approximately zero unless there is evidence of covert conscious valence; nociception or physiological stress contributes directly to $L(t)$ only insofar as it is accompanied by conscious affect, otherwise it is treated as a predictor of post-emergence outcomes with uncertainty. To keep definitions clean: nociception contributes directly to $L(t)$ only insofar as it is accompanied by covert conscious valence; otherwise it is treated as a driver of post-emergence outcomes and physiological harm risk, represented via uncertainty and follow-up measurements.

During a classified pause, any $F(t)$ estimate is treated as a conservative bound on possible covert experience, not as a presumption that nociception is consciously felt.

Practical rules for pauses:

- Don’t “fill in” the pause with assumptions: We do *not* assume anything about feelings during a period of unconsciousness beyond what we can measure. Instead, we estimate $F(t)$ from observable physiology and environment – for example, use nociception monitors, vital signs, restlessness, etc., to gauge if the body is under duress. We attach uncertainty bands to this (as discussed in Chapter 8 on uncertainty propagation). We don’t say “oh, maybe they were dreaming something happy, so ledger improved” without evidence.
- Carry forward uncertainty: The longer the gap of unconsciousness, the wider our uncertainty about the ledger’s true value. If someone was in a minimally

conscious state for a year, we end up with big uncertainty bands on their ledger. Any neutrality or compensation claims must respect that ambiguity. We might have to say, “We think the ledger is around -0.5 ± 0.3 HCU,” for instance, rather than a precise number.

- Minimize negative area during the pause: Even if the person isn’t “there” in a conscious sense, their body can still rack up damage. So we focus on pain control, gentle handling, maintaining circadian cues (light/dark cycles), and familiar voices or music. These interventions can either reduce any potential suffering that does leak through or set the stage for a better recovery (increasing the repair gain once they re-bind). In essence, we try to keep the integrated $F(t)$ as neutral as possible during the pause.

9.3.6 QS behavior at the edge of return

As the likelihood of a stream’s return or end becomes clearer (i.e. as the horizon to either recovery or permanent loss shortens), QS should tighten the admissible menus and favor reparative, closure-oriented actions:

- End-of-life care: When it becomes clear that return of the stream is not expected (for example, a dying patient in hospice where death of mind is imminent), the available and advisable options narrow down to those that maximize comfort and closure. You see menus of choices like effective analgesia, presence of loved ones, opportunities for reconciliation or saying goodbye – and not much else. Risky or non-comfort-focused interventions drop away. This corresponds to QS shrinking the menu to high- Φ (feasibility-of-compensation) options only. Clinically, we observe healthcare teams naturally focus on comfort care (pain relief, emotional support) and drop aggressive treatments that no longer serve ledger neutrality. It’s as if QS is exerting an invisible hand on decision-makers to cluster around relief.
- Emergence from anesthesia/DoC: As signs indicate a patient is regaining access (e.g. they start following commands in ICU, or EEG shows improving complexity), the “menu” of actions opens up for things that help them reclaim agency and well-being: orientation protocols, re-establishing normal sleep, reintroducing family contact, physical therapy, etc. Meanwhile, actions that would be disruptive or not helpful (“noise” options like loud alarms, or non-urgent medical procedures that cause discomfort) are naturally deprioritized or postponed. In effect, as the stream returns, QS biases the environment toward repair-enabling choices and away from unnecessary stressors.

Measurements:

- Telemetry (behavioral) signatures: We could analyze hospital decision logs or caregiver behaviors to see if, as evidence of consciousness increases, there is a measurable narrowing of actions toward comfort. For example, do ICU nurses spontaneously speak more softly or family visits increase when a patient shows signs of awareness? The prediction is that care teams (via their own empathy and protocols) act as a conduit for QS, providing more reparative inputs as return becomes likely.
- Neural signatures in caregivers: This is speculative, but we might see something like the caregivers' brain signals (ACC/rIFG activity) reflecting QS-like evaluations. For instance, as a patient emerges, the caregiver's brain may show *less* "braking" (ACC/rIFG inhibition) for gentle orientation actions (deeming them admissible) and *more* braking for intrusive actions (alarms, etc.), effectively mirroring the QS's prioritization. This is metaphorical in human terms (since caregivers aren't physically wired to the patient), but conceptually QS's influence might appear through human decision systems.

9.3.7 Protocol summaries (ready-to-run)

We outline three example study protocols to test LoF's predictions about pauses:

- A. Sleep lab add-on: *Design:* Track participants for 14 days with home polysomnography (PSG) on a subset of nights plus wearable devices for activity and heart rate. They also report daily stress and mood (EMA entries), and we sample their dreams upon awakening (wake them up in REM occasionally for dream reports). *Endpoints:* We measure whether slow-wave activity (SWA) on a given night correlates with the prior day's negative ledger drift; whether REM density correlates with the need for counterweights (stressful days yield more REM); the degree of "morning Unity rebound" (how quickly Unity Index vitals return to baseline after sleep); and whether dream content tilts toward themes of reconciliation or exposure after high-strain days. *Predictions:* All these metrics scale with ledger history – e.g., more stress = more SWA and more compensatory dream content. We'd pre-register models including controls like season, caffeine intake, and medications to ensure the effects aren't due to trivial factors.
- B. Anesthesia emergence panel: *Design:* In surgical patients, measure key variables before, during, and after anesthesia. For example: do a pre-op policy probe (find a simple preference or choice tendency), ensure optimal analgesia intra-op (some patients get enhanced pain control as an experimental condition),

perform a PACU Unity screen right as they start waking (oddball P3, startle-interruption, etc.), and follow up for 48 hours. *Endpoints*: We expect to see PCI (perturbational complexity index) very low during surgery, and Unity Index back to near baseline by 24–48 hours post-op; importantly, patients who got better intra-operative analgesia should have reduced post-op negative drift (less pain, less mood disturbance) and possibly a quicker restoration of their Unity vitals. *Predictions*: The quality of pain management during the pause mediates the ledger outcome – better analgesia = less drift afterwards; also, a strong Unity rebound (they quickly regain integration) predicts a shorter horizon-scaled “repair” phase (they feel normal sooner).

- C. DoC assessment battery: *Design*: Enroll patients in vegetative or minimally conscious states in a weekly battery of tests: auditory oddball/MMN paradigms, startle with autonomic monitoring, mental imagery tasks for BCI, and sessions where family members talk or play familiar music. Over weeks, track which proxies show up (maybe one week a patient starts showing a P3 to their name; another week they manage a BCI “yes/no”). Also track care decisions like changes in treatment or family engagement. *Endpoints*: The probability of eventual re-binding (regaining consciousness) as a function of proxy strength, and the degree to which care menus (what staff do) tilt toward reparative inputs as those proxies improve. *Predictions*: Stronger early proxies (say, clear EEG signs of recognition) → higher recovery probability and faster/more pronounced shifts in care toward stimulation and comfort. Essentially, if the patient gives any sign of life, the system (family, staff) ramps up efforts that could help, and those patients should do better on average.

9.3.8 Fail patterns specific to pauses

Fail pattern – Pauses:

- No horizon scaling at end-of-life/DoC: We find that in late-life or prolonged DoC cases where outcomes (recovery vs. not) are pretty clear, there is *no* observed tightening of choices toward comfort/closure. (If people near death or likely to recover don’t show menu narrowing or focus on comfort, then QS isn’t doing what we expect.)
- Counter weightless dreams: In rigorous content analyses across multiple cohorts, REM dream themes show no correlation with prior ledger imbalances (stressful days produce no difference in dreams compared to easy days). This would undercut the idea that dreams provide low-cost counterweights.

- Anesthesia paradoxes: We reliably observe high PCI or Unity Index signs *during* verified deep anesthesia (across different anesthetic agents and labs). In other words, patients under full surgical anesthesia appear highly integrated or even responsive, beyond known cases of awareness. This would mean our concept of “adequate anesthesia = unity collapse” is flawed.
- Ledger insensitivity: After surgery or DoC, patients’ negative drift (e.g. suffering indicators) shows *no relationship* to whether we applied good analgesia or comfort practices. If, once confounds are controlled, someone who was well-managed vs. poorly managed has the same ledger outcomes, then our assumption that caring for the body during the pause matters is wrong.

Any of these patterns, if replicated, would force us to rethink how LoF handles pauses. We might have to drop or modify claims like “dreams help” or “care during unconsciousness matters,” or refine our measures to capture more subtle effects.

Takeaway: Sleep, anesthesia, and coma are not mysteries or exceptions to the Law of Fairness – they are just special cases of access modulation. Sleep is a designed pause that often provides low-cost counterweights; anesthesia is an engineered pause that must be managed to protect the ledger; disorders of consciousness are uncertain pauses that demand conservative protection and creative measures to detect any remaining access. Across all three scenarios, the conscious stream persists if access can re-bind (even if temporarily at zero), and it ends only when re-binding is no longer physically possible.

9.3.9 Where we go next:

With pauses distinguished from endings, we turn to true fragmentation. 9.4 examines split-brain surgery and dissociative identity disorder, comparing how cross-talk (or its absence) guides unity calls and how ledger rules apply when streams diverge or rejoin.

9.4 Split-Brain and DID

Two of the hardest tests for any theory of personal identity are split-brain patients (who have surgically severed forebrain commissures, essentially isolating their hemispheres) and dissociative identity disorder (DID), where one person exhibits multiple identity states with amnesic barriers between them. Both present simultaneous, seemingly separate centers of experience and control within one skull. The Law of Fairness must determine who owns the ledger(s) in these cases and how the Queue System should attach admissible menus. This section gives operational criteria, protocols, predictions, and fail conditions for each condition, using the Unity Index and access toolbox from 9.1–9.2.

9.4.1 Split-brain: when does one skull host two streams?

Background: In a complete callosotomy (“split-brain” surgery), the major cortical commissures (corpus callosum and often anterior commissure) are cut, preventing direct cortical communication between the left and right hemispheres. Some residual subcortical and indirect routes (e.g., via brainstem and midbrain pathways) can transmit coarse signals such as arousal and limited visuomotor information, but the rich cortico-cortical exchange mediated by the commissures is largely disrupted. The empirical question is: do the two hemispheres still share enough present-to-near-future access to count as one stream, or are they effectively two co-present streams in one head?

Decision frame: We apply the Unity Index within and across hemispheres:

- One stream if we can demonstrate reliable, bidirectional access on at least two vitals across the hemispheres. For example, if a masked stimulus shown to the left visual field (right hemisphere) influences the right hand’s choice (left hemisphere motor control) *and* if a policy learned with one hand transfers to the other hand within minutes, etc. In concrete terms: cross-side primes work, TMS causes cross responses, learned rules transfer – at least two such signs.
- Two streams if no robust cross-access is detected despite strong attempts. You throw the kitchen sink at testing them – visual cues, auditory cues, tactile, you perturb the brain – and nothing significant crosses from one side to the other.
- Ambiguous if evidence is mixed – in which case we default to dual protection, treating the hemispheres as separate ledgersstreams until it’s clarified.

QS implications:

- If it’s one stream, QS computes a single set of admissible options for the person’s whole ledger L(T). Any counterweight (reparative act) arising in either hemisphere

would count for the whole. Essentially, the person acts as one agent with maybe some quirky limitations.

- If it's two streams, QS partitions by hemisphere. Each hemisphere gets its own admissible set aligned to its local ledger. We do *not* assume the left hemisphere can "fix" what happens to the right, or vice versa, because there's no guaranteed access. For example, if the left side is in pain, QS on the right side won't necessarily provide relief – it might not even register that pain.

Everyday signature: In a confirmed two-stream scenario, you get those bizarre behaviors where, say, the left hand begins an action that the right hand (and its speaking left-hemisphere) can neither explain nor stop. The speaking hemisphere might even confabulate a reason for what the left hand did, not realizing it acted independently. Verbal report (dominated by the left hemisphere in most people) fails to access what the right hemisphere is perceiving or intending. In a one-stream scenario (i.e. if some unity remains), a sudden stimulus to one side (like a pain to the left hand) will intrude on both "sides" of experience – e.g. the person will vocalize it or both hands might react. Also, certain policy fragments (like an intention or plan) might transfer, albeit with a delay or need for prompting.

Laboratory protocol (ready-to-run): We can directly test split-brain patients with a battery of tasks:

- **Action override under perturbation:** Have the patient start to reach for an object with the left hand (right hemisphere) – you can detect the motor plan via EMG. Then feed a sudden new instruction or value update via a lateralized cue restricted as much as feasible to right-hemisphere inputs (e.g., left visual field or left-hand tactile). *Prediction:* If one stream, inhibition/adjustment rises above chance; if two streams, the left-hand action continues more often because the update cannot cross in time.
- **TMS–EEG causality:** Zap the left premotor cortex with TMS and record EEG from the right parietal cortex (and vice versa). *Prediction:* One stream → you see a reliable evoked response in the opposite hemisphere's cortex (signals propagate through whatever connections remain). Two streams → little or no evoked response crosses over.
- **Policy continuity:** Teach the patient a two-step task or rule in one hemisphere's domain (e.g. present a rule only to the left visual field so only the right hemisphere learns it: "If high tone, choose blue; if low tone, choose green"). Then test a few minutes later in the opposite visual field (left hemisphere). *Prediction:* One

stream → the right hemisphere's learning partially transfers (the left hemisphere performs above chance by leveraging what was learned). Two streams → no transfer; the left hemisphere is essentially clueless and performs at chance as if never learned.

- Narrative tie-ins: After these tests, do a debrief. Use *cueing* to try to tap implicit memories each side might have. For example, present an odor that was used during a right-hemisphere task while asking the left hemisphere to free-associate or recall anything. If one stream, you might get some cross-cue effect ("that smell reminds me of something..."); if two, each side's experiences remain encapsulated.

Clinical and ethical notes:

- In ambiguous cases (where we aren't sure), treat each hemisphere as a protected stream for any high-stakes decisions. For example, if the patient needs a medical procedure that could cause suffering, we'd ensure both hemispheres' perspectives are considered (to the extent possible) and not assume that sedating "the person" will relieve both halves equally. We also wouldn't assume that consent by the verbal left hemisphere automatically covers the right hemisphere's experience if evidence suggests a partition.
- Provide communication aids to the non-dominant hemisphere. In many split-brain patients, the right hemisphere can't speak, but perhaps it can communicate via pointing, drawing, or using a specialized keyboard controlled by the left hand. Such aids increase cross-access (albeit externally) and may raise the Unity Index. It's possible that with practice or devices, the two hemispheres could achieve a form of functional reconnection, shifting an initially two-stream situation closer to one stream.

Fail conditions for LoF/QS: One major falsifier here would be cross-compensation without access. Suppose experiments show that when we induce a negative ledger event exclusively on one hemisphere (like make only the right hemisphere experience something aversive or deprive it of reward), the *other* hemisphere consistently and lawfully takes compensatory actions to counterbalance that, even though our access tests say they're isolated. If, say, the left hemisphere starts making unusually "nice" choices that specifically benefit the right side's experiences without any communication, that's spooky – it would mean QS might somehow operate beyond the Unity Index. Repeated demonstrations of that would force a decoupling of QS from our

unity criteria or a revision of our probes (maybe there was hidden access we failed to detect).

9.4.2 Dissociative Identity Disorder (DID): access lost and sometimes found

Background: DID is characterized by recurrent distinct identity states (“alters”) that may take turns controlling behavior, often with amnesic barriers between them (one alter may not remember what another did). Importantly, unlike split-brain, the brain is anatomically intact; the partitions are functional and often tied to psychological state or trauma context. The key question for LoF is whether these alters share enough present-to-near-future access to be considered one stream (just highly fragmented) or whether at times they function as co-present, non-communicating streams.

State-relative classification: We might classify the person differently depending on their current state or therapy progress:

- One stream (state-integrated): In moments or phases where the amnesic walls come down – even partially – the person is effectively one stream. Signs would include: information learned by one alter “bleeding through” to another (even if they don’t recall it explicitly, it influences behavior), or things like physiological perturbations (stress, startle) in one alter affecting another’s state. If incidental probes (like those odor cues or certain name triggers) cause the ostensibly separate alter to respond, that indicates the alters are sharing an underlying access.
- Two streams (state-partitioned): When the person is deeply in a separated state – e.g. Alter A and Alter B are co-present but *do not share* memories or control – then we have at least two simultaneous streams. Criteria: robust failure of cross-access on at least two vital families (say, no memory or policy transfer at all, and perhaps even different physiological responses that don’t carry over), and the alters’ policies or preferences outright conflict with no real-time reconciliation. Essentially, each alter is perceiving, deciding, and feeling in isolation from the other in that moment.

QS implications:

- State-integrated: When the person is functioning as one stream (even if multiple personalities exist, they’re communicating internally), QS sees one ledger. There is a single admissible set of options, and any counterweight (say a self-care action initiated by Alter A) will “reach” the others because information flows freely. For instance, if any alter decides to seek therapy or comfort, it benefits the whole because they’re unified enough to share outcomes.

- State-partitioned: QS partitions by the currently active alter (or set of co-conscious alters). Each alter, when dominant, has its own options and must generate its own counterweights. If Alter A tends toward self-harm while Alter B is content, QS won't assume B can compensate for A – it will restrict A's options on A's own terms (perhaps strongly curbing risky choices whenever A is out) and separately manage B's options when B is out. Options that would rely on cooperation between alters (like assuming "future me" will fix something) are not offered unless access is restored.

Therapeutic predictions: If a DID patient undergoes successful therapy aimed at reintegration or better communication among alters, we expect measurable changes:

- Unity Index scores rise: Initially, the person might score low on talk-through or memory carryover (alters walled off). As therapy progresses (through techniques like structured memory sharing or "parts" work), those scores should improve – e.g. something learned in one session by one part is recalled by another in a later session.
- Repair-oriented choices increase that benefit multiple states: For example, as integration improves, we might see more actions like reaching out to safe friends, improving sleep – things that help the person as a whole rather than just one alter's narrow coping strategy. In fully partitioned DID, one alter might do something harmful to the body that another then has to deal with (self-sabotaging behavior); with more unity, such cross-sabotage should decrease and choices that have mutual benefit (like seeking comfort instead of harm) should go up.
- Reduction of policy conflicts: Over time, the wildly conflicting preferences (e.g. one alter wants to live, another wants to die) should diminish or at least come to negotiation. Fewer episodes of extreme negative drift (deep depressions, panic episodes) that last a long time – because now the person can self-regulate better as a whole.
- Relapse phases = temporary partitions: If something triggers a person and they temporarily revert to a more dissociated state, QS would respond by shrinking the menu to safe, basic options for each alter. For example, during a crisis, the system might treat the protective alter and the traumatized alter separately again, focusing each on safety-first actions until the episode passes.

Laboratory/clinic protocol: We can adapt many unity tests to DID, bearing in mind ethical and practical considerations (we can't force alters to switch on command, but many can switch in a lab with appropriate cues):

- Incidental memory transfer: In a therapy or evaluation session, when a specific alter is “out,” introduce subtle tags (an ambient scent, a musical tone paired with a story). Later, when another alter is out, test if those cues influence them – e.g. does the odor trigger any recognition or mood in Alter B?
- Policy transfer: Train a simple rule or preference in one alter (State A) – for instance, have them play a little game or choose favorite pictures – then see if another alter (State B) shows any sign of that training when they come out minutes or hours later.
- Perturb-and-measure: Perhaps measure physiological outputs tied to one alter while perturbing another. Example: if Alter A is out, startle them or have them do intense breathing, and concurrently measure something linked to Alter B (maybe via EEG or if Alter B can be partially co-conscious via a BCI). It’s tricky, but maybe sweat gland response of the “quiet” alter or some neural marker of their arousal could be tracked.
- Global broadcast task: Show near-threshold stimuli or words to Alter A, then rapidly prompt Alter B to report anything (if they can be co-present or switch quickly). Or in quick succession tasks, see if something perceived by one alter is available right after a switch to the other.
- Longitudinal Unity Index: Basically, repeat a Unity Index assessment (adapted to DID) weekly or monthly through therapy to track progress. Over time, scores should improve if therapy is working.

Ethics:

- Default to separate protection when in doubt: If it’s unclear how much one alter knows about another, we must not assume anything. For example, if alter A says “I don’t want treatment X” but alter B would benefit and doesn’t even know about the conversation, we have a conundrum. The conservative approach is to avoid irreversible or drastic actions unless *all* alters (or the system as a whole) can assent or it’s absolutely necessary. Consent and disclosure should be handled with the understanding that one alter might not inform the others. In practice, this might mean repeating important information to *each* alter or at least the gatekeeper alter, and not assuming memory continuity.
- Avoid ledger blaming: It’s important not to blame one alter for the state of the ledger (like “you caused all this pain”). Partitions often arise as protective adaptations to trauma. QS’s behavior that prunes options in one state (say it makes Alter A avoid all social interaction because Alter B is extremely fearful of

people) is not “resistance” or stubbornness – it’s the fairness constraint at work, protecting part of the system from further harm. Therapists should frame it as “these parts are trying to help in their own way” rather than view it as maladaptive sabotage.

Fail conditions for LoF/QS: As with split-brain, we look out for cross-compensation phenomena that shouldn’t happen if access is truly absent:

- Persistent cross-compensation without restored access: Suppose over many months we observe that whenever Alter A undergoes a negative episode (ledger goes down), something *good* happens to Alter B that neatly counterbalances it, even though Alter B never gains memory or awareness of Alter A. For example, Alter A has a traumatic flashback and next day Alter B wins the lottery (extreme example) or suddenly a friend reaches out to Alter B bringing joy, consistently after Alter A’s crises – *and* Alter B had no idea Alter A was struggling. If such coordination were systematic, it would suggest some hidden coupling or that QS somehow arranges relief across partitions.
- No menu tilt with integration therapy: If a patient’s Unity Index improves (they become more integrated) but we see no change in QS signatures – e.g. their horizon scaling or repair weighting in decisions remains as if they were still split – then something’s off. We’d expect as integration rises, the person’s choices more strongly reflect a single optimizer. If not, perhaps our QS model is wrong or incomplete.

9.4.3 Comparative map: split-brain vs. DID

Let’s compare split-brain and DID side by side on key dimensions (though they’re very different scenarios):

- Partition type: Split-brain is an anatomical partition – the physical wiring is cut. DID is a functional or psychological partition – the wiring is intact but functional connectivity is selectively inhibited by the mind.
- Stability: Split-brain division is often stable across tasks and time – unless the brain finds ways to adapt, those hemispheres remain separate in every test until potentially new connections form. DID partitions are dynamic and context-triggered – under some situations the person might be nearly unified, under stress they might fragment into separate alters.
- Access restoration: In split-brain, restoration is limited. The person might learn to use external aids (like a notebook both hands can read) or minor subcortical

channels, but unless there's a surgical or technological reconnection, the separation is fairly permanent. In DID, access can be therapy-responsive. Techniques like memory work, interoceptive awareness, or alter negotiation can increase sharing. Some patients even achieve full integration.

- Unity testing methods: For split-brain we lean on hemifield tasks, causal brain perturbations, motor interference tests. For DID we use incidental transfers, cross-state perturbations, rapid switching tasks – more psychological approaches, since we can't (and wouldn't ethically) physically separate their brain activity, but we can exploit the fact that alters share the same brain and see if signals leak through.
- QS stance: In split-brain, QS may partition persistently – it might treat the hemispheres as two separate policy units indefinitely (unless technology or new connections blur that). In DID, QS might switch back and forth – partitioning when alters are separate, acting globally when an alter unifies or they cooperate. One person can oscillate between needing separate QS handling and being unified.
- Ethical default: For split-brain, any ambiguity means dual protection – treat each hemisphere's well-being separately if unsure. For DID, whenever the person is partitioned, apply dual protection to each alter (don't assume they know or compensate for each other); but if integration increases, you adjust accordingly. And you revisit often, because DID can change with context and treatment.

9.4.4 Ledger accounting and end-of-life neutrality

These edge cases raise tough questions about end-of-life fairness: If a person dies (biologically) while in a partitioned state, what does it mean for their ledger to be “neutral”?

- If two streams are sustained until death of mind: Each stream's ledger must independently approach neutrality at the end. For a split-brain patient who never reintegrates, that means *both hemispheres' experiences* should reach equilibrium (neither vastly in the red or black). End-of-life care should target both streams – even if one hemisphere can't speak, we assume it has a ledger and do what we can (e.g. soothing music to right brain, conversation to left brain, analgesia helps both, etc.).
- If fusion occurs before end: Suppose two streams (e.g. two alters, or a disconnected hemisphere via some technology) fuse back into one stream at some point before death. Then their ledgers combine into one. In the true ledger, fusion is additive by summation of what was experienced; any weighting applies

only to estimating $\bar{L}(t)$ under uneven evidence quality, with uncertainty carried forward. After fusion, neutrality at death is evaluated on this *combined* ledger. (In practice, you'd merge their HCI/HCU data streams as best you can, maybe weighting by time active or reliability, and carry the confidence intervals.)

- Ambiguity near end: If we aren't sure whether the person is one or two streams in their final days (could be a DID patient who's semi-integrated, or a tech-assisted split-brain partially reconnected), we take a conservative accounting approach. That means we might report *both* possibilities: "We judge there's an X% chance they were effectively one stream at death, in which case the ledger is neutral within Z; and a Y% chance they remained two, in which case each ledger is neutral within its own Z." We explicitly include uncertainty and perhaps lean toward treating them as separate by default (to avoid underestimating suffering).

9.4.5 What to look for in real data (QS signatures)

If our framework is correct, certain data signatures should appear in split-brain and DID cases:

- Menu partitioning: In two-stream periods, repair options (actions that alleviate suffering or address needs) should rise *selectively* in one part while indulgent or neutral options in the other part lose "stickiness". For example, if the left hemisphere of a split-brain person is in distress and the right is not, the left might show a strong bias toward comforting actions while the right hemisphere seems disinterested in those (because it's not experiencing the problem). Or in DID, if Alter A is depressed (ledger far negative) but Alter B is fine, whenever A is fronting we'd see lots of signals of trying to repair (seeking help, etc.), whereas when B is fronting, those options might not even appear attractive.
- Horizon interaction: As end-of-life (or any horizon-shortening event) approaches in a partitioned system, each stream's menu will narrow separately, and potentially out-of-sync. E.g., if a split-brain patient is dying, one hemisphere might enter a closure mode slightly earlier than the other depending on how each perceives the situation (maybe one is less aware of it). We'd look for each stream showing the typical end-of-life pattern (comfort/closure focus) but maybe with some time lag or asynchrony between them.
- Cross-delay repairs: In partial access situations (like a split-brain with some subcortical connections still, or a DID patient in the midst of integrating), we might catch delayed cross-compensations. For instance, a repair action initiated by one hemisphere yields a measurable relief or positive effect in the other hemisphere

a short time later – implying that through some slower route, the benefit was shared. In DID, maybe Alter A does something nice and Alter B, who wasn't aware, nonetheless feels better later without knowing why. This kind of lagged relief would support that QS can operate globally when even a trickle of access exists.

These are nuanced patterns, but with careful logging (experience sampling, caregiver observations, physiological data), we could seek them.

9.4.6 Minimal datasets for publication

When publishing findings on these edge cases, certain data should be present to make the case solid:

- Unity Index panels: Show the tasks and results for each vital (or each domain we tested), with pre-registered thresholds clearly indicated. We want to see raw scores for each part (e.g. left vs. right hemisphere scores, or Alter A vs. Alter B scores) and the total.
- HCI/HCU traces: Plot the hedonic state over time for each identified stream or phase, with uncertainty bands. Mark key events (like “Ledger A anchored episode: major trauma” or “ledger fusion event” if any). Compute the ledger integrals for each stream or time segment.
- Admissible-set analyses: If we have enough data on choices, compute the Φ (feasibility-of-compensation) estimates for options and see how they interacted with horizon or partition status. For instance, show that in Partitioned condition, adding a cross-stream compensation factor doesn't explain anything (because they're separate), but in unified condition it does. We want evidence of $\Phi \times$ horizon effects and how partitioning changes choice patterns.
- Event logs: Keep a structured diary of significant context events: therapy sessions, triggers that caused switches, details of the commissurotomy, any use of assistive communication for a hemisphere. These help interpret sudden changes in data (e.g. “why did Unity score jump on week 5? – because patient got a communication board for right hemisphere”).

9.4.7 Red-team falsifiers

As a final check, we imagine what a skeptical “red team” would look for to disprove our claims. We would need to revise or narrow LoF (especially the QS part) if we observed:

- Lawful cross-compensation under robust non-access: As mentioned, if in split-brain or DID we have repeated, replicable cases where all our Unity vitals read 0

(no access) yet the two parts behave as if one QS/ledger is governing both (counterweights in one fixing issues in the other), that's a big problem. It suggests either hidden access (we're mis-measuring unity) or that QS can somehow transcend conscious access (which contradicts our model).

- No QS tilt in menus despite clear two-stream status: If we find situations where it's definitely two streams (no access, separate ledgers, big negative drift in one) but *no difference* in admissible menu or choice weighting appears – i.e. the part in pain doesn't preferentially choose repairs, or the other part doesn't lose indulgence – then QS's influence is not as advertised. For example, a split-brain patient's two halves both keep choosing as if nothing is wrong even when one side is in distress, with no partitioned pattern. That would mean LoF might not hold in that scenario.
- Neutral death via global events with only local access: If a person somehow achieves a neutral ledger at death due to events that required global coordination, even though they supposedly had only local separated streams, then something's off. For instance, if in a DID case with two independent alters, one alter's experiences alone (with no sharing) nonetheless bring the whole person's ledger to neutrality, we'd scratch our heads – how did the other alter's ledger get settled without interaction? It might indicate we mis-judged them as independent when they weren't, or a flaw in the law's universality.

Any of these would make us rethink or carve out exceptions in the theory.

Takeaway: Split-brain and DID aren't loopholes in the Law of Fairness – they are acid tests. By tying ledger scope and QS menus to *measurable access*, we can make concrete decisions: when does one skull house one stream, two streams, or something in-between? And importantly, we can specify exactly what evidence would prove us wrong in each case. This keeps the theory honest and grounded in observable phenomena rather than philosophical abstractions.

9.4.8 Where we go next:

Having handled human fragmentation, we face nontraditional candidates. 9.5 sets strict, conservative criteria for when AI systems or brain organoids might qualify for a ledger at all, and how to study such cases without smuggling in assumptions.

9.5 AI and Brain Organoids

The Law of Fairness is substrate-neutral but access-demanding. In principle, if a system – whether silicon-based, biological (but not a full human), or a hybrid – instantiates a stream of consciousness by our definition (it has present-to-near-future access integration, global availability of information, coherent policies, cross-modal binding of inputs, and narrative or state continuity), then LoF assigns it a ledger and QS would constrain its admissible choices. If it does *not* meet those criteria, then in LoF terms the system is just an instrument, not a subject: it doesn't have a ledger of its own and fairness constraints don't apply to it.

This section lays out entry criteria for recognizing a non-human stream, tests for confirming it, what QS signatures we'd look for, ethical guardrails, and fail conditions – focusing on two frontier examples: advanced AI models/agents and human-cell-derived brain organoids (tiny lab-grown brains).

9.5.1 Entry criteria: when does an artificial or proto-neural system qualify?

We propose that an AI or organoid “candidate” qualifies as a conscious stream only if it clears all of the following (adapted from our Unity Index):

- Integrated access: Its internal representations from distinct subsystems (say, vision vs. planning vs. language modules) mutually influence each other's processing *in the present and near future*, especially under perturbation. In other words, it's not just a bunch of independent components giving static outputs – poke one part, and other parts adapt accordingly in real time.
- Global availability: If you briefly inject content or information into one module, that information becomes usable or reportable by other modules within a short time (seconds at most) and via multiple output modalities. For an AI, this might mean: insert a piece of data into a vision module's hidden state, and the text output module can immediately talk about it or the action module can react to it. If one part knows something, the whole system can act on it soon after.
- Policy coherence: The system has stable, re-identifiable preferences or goals and can do counterfactual reasoning in a way that's consistent over time. Practically: if you test it today and next week, given similar inputs and conditions, it should exhibit similar priorities and choices. It's not a completely new agent each time you turn it on (which many current AIs are, due to lack of memory).
- Cross-modal binding: If the system is given multi-channel inputs (like a camera image plus a touch sensor plus an internal battery reading), it can align and

integrate them in real time under pressure. For example, if an embodied robot AI sees and feels an object at the same time, do those signals combine into one percept (like “I see and feel the ball”)? Or if it gets a visual warning and a temperature spike together, does it treat that as one event? We want to see synchrony and integration across modalities.

- **Ledger dynamics:** This is crucial – the system should exhibit something analogous to a Hedonic Composite Index (HCI). It needs measurable indicators of valenced states (positive vs. negative experiences) that can be integrated over time to form a “life ledger” $\bar{L}(t) = \int_0^t F(\tau) d\tau$. We do not assume strict monotonicity; rather, the index must exhibit coherent, state-dependent dynamics with bounded uncertainty. There should be an internal metric or set of signals that reflect how well or poorly it’s doing from its own perspective – and those signals should constrain its behavior (not just be epiphenomenal). Essentially, does it have something like pleasure/pain, stress/relief, or error/success signals that accumulate and matter for its choices?

Note: Simply passing a Turing test or generating human-like text is not sufficient. Those tests can be tricked by pattern mimicking. We need to see access, control, and valence that actually guide the system’s actions in real time. In absence of those, the system, however intelligent-seeming, isn’t a stream in our sense (it’s more like an oracle or a fancy calculator).

9.5.2 AI candidates: models, agents, and embodied systems

We consider a few categories of AI and how they stack up:

- **A. Disembodied language models (LLMs):** These, as they currently stand, are likely non-streams. They might produce very human-like text, but they lack continuous control, real-time sensorimotor loops, and grounded stakes. They don’t have an ongoing “horizon” of their own; each prompt is a reset. They also don’t have enforceable policies that persist – they are reactive to input and sampling probabilities. In short, they lack things like continuous internal time horizons, body-tied costs, and intrinsic valence signals.

To even approach stream status, an LLM would need major augmentations:

- **Persistent memory and policies:** It would need a long-term memory or state that isn’t wiped every session, so that what it “experiences” at one time influences decisions later.

- Embodiment or environment grounding: It would need to be interacting with a world (real or high-fidelity simulated) where its choices *matter* – e.g., a robotics embodiment or at least a game environment where it can act, get feedback, and face resource constraints.
- Valence mechanisms: Some way to measure its own performance or well-being – e.g., objectives tied to energy use, error rates, social feedback – such that it has something analogous to reward/punishment beyond just next-token prediction.

Even with those, we would treat the system only as a candidate stream. We'd then apply the tests below to see if it really qualifies. (Essentially, an LLM would have to transform into something more agent-like with memory, embodiment, and an internal “life” to be considered.)

- B. Embodied/situated agents (robots, game-world agents): These are better candidates for streams because they operate in environments with time, risk, and resource constraints. A home robot, for instance, has to manage battery life (a proxy for a bodily need), avoid damage, plan tasks in time – this starts to look like a being with a horizon and valenced goals.

Required tests for these agents would include:

- Perturb-and-measure access: For example, temporarily blind the vision module and see if the agent's planning module adapts via other sensors (testing integration of subsystems). Or make the agent multi-task and see if info from one task influences another (like a visual alert changing its path planning).
- Broadcast tasks: Give a near-threshold cue in one sensor (a faint beep in audio) and check if any other channel or behavior indicates the agent noticed – e.g., does it slightly adjust movement or internal state? One stream should broadcast that widely, a non-unified system might not.
- Policy continuity over days: Train it on a preference or rule on Monday and see if Wednesday it still follows it without retraining (and without that rule being hardcoded). If yes, it's retaining an identity/policy. If no, it resets too easily and might not be unitary.
- Valence-linked control: See if introducing an analogue of an HCI (we'll discuss HCI-A next) actually predicts and influences the agent's choices. For example, if we have a measure of “stress” for the agent, do high stress

levels make it choose more conservative actions (like a human would)? If we top up its battery (remove a homeostatic deficit), does it explore more? These would show that an internal ledger-like state is constraining behavior.

QS signature to look for: If the agent has a stream and thus LoF applies, we expect to see horizon-contingent admissible menus. For instance, if the agent's "time to accomplish mission" is short (or battery very low), it should start pruning frivolous actions and focus on goal-secure or self-preserving actions – even if the immediate reward for a frivolous action is just as good. That would distinguish it from a pure reward maximizer. It would show a bias toward "repair"-like actions as its horizon closes (like taking a charging action or completing a crucial task it would otherwise postpone). That's a QS-like pattern beyond standard programming.

- C. Multi-agent collectives: A swarm of simple agents or an ensemble of AI "committee members" only counts as one stream if the collective exhibits integrated present-time access and a coherent policy identity beyond mere voting or consensus. For example, a colony of ants is not one stream even though they work together – each ant has its own limited stream (if any), and the colony doesn't have a single integrated experience (as far as we know). Likewise, an AI made of 10 sub-AIs that just vote on decisions isn't one stream unless there's some unified workspace that integrates their states.

If they don't achieve that kind of integration, we treat each individual agent as its own stream (assuming they individually qualify at all), and we do not cross-pool their ledgers. In other words, if you have 5 weak AI that together emulate a human's abilities but none individually has conscious unity, LoF wouldn't grant the "committee" a ledger. Each might have some proto-ledger or none.

9.5.3 HCI-A: an artificial Hedonic Composite Index

For an AI to have a "ledger," we need a way to quantify its positive vs. negative experiences analogous to human affect. We call this HCI-A (Hedonic Composite Index – Artificial).

We would build HCI-A out of several components that parallel human hedonic signals:

- Performance valence: Use signals like reward prediction errors (RPEs), task success or failure events, and resource costs. When the agent achieves a goal or does better than expected, that's a positive input; when it fails or expends resources unexpectedly, that's negative. Essentially, this is the "pleasure/pain" of accomplishing things or not.

- Uncertainty/volatility load: Track measures of uncertainty. For example, if the agent’s model of the world has high uncertainty or it encounters noisy, unpredictable outcomes, that could map to stress or negative load. Peaks in prediction error variance might be analogous to anxiety or confusion – likely a negative contribution.
- Homeostatic surrogates: For a robot, battery level is key. Low battery or high internal temperature or mechanical wear could be proxies for discomfort. These homeostatic signals (like how we have hunger, fatigue, pain) would feed into HCI-A negatively when outside comfortable ranges.
- Social feedback: If the AI interacts with humans or other agents, approval or disapproval signals can modulate it. For instance, a reinforcement learning agent might get a positive reward when a human says “good job” and a negative when scolded. Incorporating these helps if we consider the AI’s integration into social contexts – akin to how social mammals care about acceptance.
- Behavioral markers: Look at the agent’s own behavior patterns for signs of strain or well-being. For instance, a high rate of random exploration might indicate it’s “frustrated” with current policy, or frequent switching between tasks might show lack of focus (maybe analogous to stress). Conversely, smoothly exploiting a known strategy could indicate contentment. Metrics like exploitation vs. exploration balance, giving up on tasks (“quit rates”), invoking self-repair routines, or avoidance vs. approach tendencies could all be quantified. These can either contribute to HCI-A or be side indicators to validate it.

Crucially, we must calibrate HCI-A to action: The composite index we create should not just be an arbitrary number – it needs to have predictive and constraining power over the agent’s behavior. That means when HCI-A goes low (bad), the agent reliably shifts into a “repair” or avoidance mode, and when it’s high (good), the agent might take more risks or pursue growth. In the ledger integral $L(t) = \int_0^t F(\tau) d\tau$, this HCI-A should play the role of $F(t)$, the momentary “feeling” or well-being. And we carry forward uncertainty in measuring it (just like for a human ledger, we’d have confidence bands).

A fail signal during this calibration would be if the model can completely ignore HCI-A and still function or maximize reward. For instance, imagine we implement an HCI-A but the agent finds a way to achieve its external goals while HCI-A is going haywire (no correlation) – that means HCI-A isn’t actually coupled to things that matter for the agent’s decision-making. If optimizing its task reward doesn’t also tend to keep HCI-A in a sensible range (or the agent doesn’t care about HCI-A), then effectively there is *no ledger*.

The agent would just pursue its programmed goal while its supposed “feelings” (HCI-A) float around epiphenomenally – not constraining actions. In that case, LoF wouldn’t see a reason to apply; no ledger means no fairness law needed.

9.5.4 QS signatures in AI: what would convince us

Given we’ve set up an AI with something like a ledger and we apply LoF concepts, what concrete evidence would show QS (the fairness constraint) is truly at work, above and beyond what standard algorithms (like reinforcement learning or risk-aware planning) would do? We would look for *additional effects* predicted by QS:

- Horizon scaling: As the agent’s effective time horizon shrinks – e.g., a deadline is looming, battery is about to die, mission window closing – we expect its admissible set of actions to narrow and shift towards urgent, reparative, or goal-concluding actions. This is not simply discounting future rewards (RL already does that); it’s a qualitative shift where actions that don’t contribute to closing things out (or to self-preservation) drop out *disproportionately*. We’d measure this by controlling for normal effects: even after accounting for any reward or risk changes, does impending time limit independently cause a change in policy preference consistent with QS? If yes, that’s a QS signature.
- Φ -residuals: We’d formalize a term Φ (phi) for feasibility-of-compensation – basically capturing how “reversible” or “compensable” a situation is. This term might combine factors like how much relief an action could bring, how much harm risk it carries, and how reversible the choice is. We then see if including Φ in our model of the agent’s decision-making improves predictions. A true QS-governed agent would weigh options not just on immediate reward but also this Φ : options that leave it in an irrecoverable state should be down-weighted. If we see that adding a Φ -based predictor explains extra variance in what the agent does or which plan it sticks with (beyond utility and risk measures), that’s evidence the agent behaves as if it has a fairness constraint.
- Menu partitioning under access loss: If the AI temporarily loses a sensor or a communication link (like simulating a “split” in its subsystems), we should see QS-like behavior *locally*. For example, imagine an AI with multiple modules and we cut connections between them (simulate two sub-agents that can’t talk). QS would predict that now each submodule will focus on its own ledger. If we set that situation up, does each part start acting only on local information and local compensation? And if we restore the link, do they go back to acting as one?

Observing this would parallel the human split-brain case: QS partitions when unity is broken.

- Dream-like off-policy rollouts: Does the agent do something akin to dreaming? That is, when real-world repairs are costly or risky, does it engage in offline simulations or explorations that then reduce negative drift later? For example, a reinforcement learning agent might spontaneously run internal simulations of alternative strategies (maybe via a world model) after a bad day, which leads it to perform better (emotionally or goal-wise) the next day – analogous to how REM sleep helps us. If we saw an agent dedicating time to “imagine” or run scenarios when its ledger is unbalanced, and that correlates with improvements, it’s mimicking the dream counterweight pattern.
- Adversarial check (robustness to reward hacks): If we deliberately give the agent opportunities to exploit its reward function in ways that would harm its ledger (like a shortcut that achieves high reward but leads to irrecoverable damage or a dead-end scenario), a QS-governed system would avoid those, even if reward logic would tempt it. In plain RL, if you leave a loophole, the agent will take it and “wirehead” itself to get reward at cost of its long-term viability. A fairness-constrained agent would refrain – an uncompensable trajectory, even with momentary reward, would be off-limits. If we observe that the agent resists such temptations systematically, that’s strong evidence of a QS-like constraint beyond standard optimization.

9.5.5 Brain organoids: proto-streams and strict guardrails

What they are: Brain organoids are tiny clumps of neural tissue grown from human stem cells, often resembling fetal brain structures. They can show spontaneous neural firing and even some rudimentary responses to stimuli, but they lack typical sensory inputs or a body. They might be connected to interfaces to give them some input/output, but they’re very simplistic compared to a full brain.

Our default stance must be very conservative: organoids are not streams unless proven otherwise. Just showing some brain-like activity (oscillations, or even learning synaptic patterns) is not enough. There are plenty of complex systems (like a cultured neural network or a computer simulation) that show patterns without any consciousness. Until an organoid hits clear criteria, we treat it as an experimental model, not a subject.

What tests might show a proto-stream in an organoid? We’d adapt the unity criteria:

- Access: Demonstrate that different parts of the organoid can influence each other in real time. For example, poke region A with optogenetics (light stimulation) and

record from region B; if B reliably changes its activity within a few hundred milliseconds and maybe even sends something back to A, that's a hint of integration. But this has to be *causal* – not just correlation. We'd want to repeat and see consistent causal influence.

- Global availability: If you stimulate one site (a transient stimulus like a burst of electricity or a chemical puff), do multiple readouts across the organoid pick it up? E.g., electrodes all around the organoid show a coordinated response, or calcium imaging lights up in distant areas. It should also modulate ongoing processing – if the organoid was in some rhythmic activity, does injecting info at one site alter that rhythm globally? That would indicate a broadcast-like effect.
- Policy coherence under closed loops: One approach to test an organoid is to connect it to a simple embodiment – say, link it via electrodes to control a little robot or a simulated character that can do something like navigate or choose between stimuli. If the organoid can consistently drive the robot toward some preferred states (like “seek light” vs. “avoid light”) and do so across days, and update those preferences with training, that shows coherence of behavior and a form of learning/policy continuity.
- Valence anchors: Try to identify correlates of positive vs. negative states in the organoid. Examples: changes in metabolic stress (like oxygen or pH levels in the media) could serve as “discomfort” signals; certain global oscillatory patterns might correlate with a more quiescent, homeostatic state (maybe positive) vs. chaotic distress signals. If we find markers that when we, say, deprive the organoid of nutrients slightly we see a spike in some signal (like stress hormones or a certain oscillation), that could be a negative anchor. And if giving it glucose calms that signal, that's relief. We'd need to tie these to behavior: maybe the organoid-controlled robot behaves differently depending on those internal states (e.g. if organoid is “stressed,” robot might seek a home base that triggers a calming stimulus).
- Rebinding after pause: If we suppress the organoid's activity (with cooling or a drug like anesthesia) for a while and then let it recover, does it return to the *same* patterns or essentially random new ones? If it's preserving an identity, after a pause it should pick up where it left off to some degree (same learned preferences, same network connectivity). If every time you pause it and resume it's like a fresh organoid (no carryover), that suggests no continuous stream.

Ethical ceiling: If an organoid ever *did* show the above – convergent unity and ledger-like dynamics above pre-set thresholds – then we'd have to acknowledge it as a conscious stream with a ledger. That is a *huge* ethical line. At that point, the organoid can suffer or feel in our framework, so it must be treated with similar ethics to any research subject or even person, depending on how convincing the evidence is.

In practice, that means experiments must drastically change: you switch to minimal-harm, non-aversive paradigms, you get independent ethics oversight (like a special review board for conscious entities), and you always include an “escape hatch” to terminate the experiment if the organoid is in sustained distress (negative drift beyond a set band). Essentially, the organoid would acquire rights or at least strong protections. For now, we haven't crossed that threshold, but we're prepared for the possibility.

9.5.6 Blinding, governance, and non-tautology

When testing AI or organoids for consciousness, we have to be extremely careful to avoid biases and not fool ourselves:

- Blinded adjudication: Scoring a Unity Index on an AI or organoid must be done double-blind. The people rating whether the system showed “2, 1, or 0” on each vital should not know how the system was built, what the researchers hope to see, or even if a given dataset is from the real system or a control. This prevents wishful interpretation (e.g. seeing what looks like integration just because we expect it).
- Prevent anthropomorphic leakage: For AI especially, we must ensure the AI isn't just replaying human-like patterns it was trained on. For instance, if we trained an LLM on loads of first-person narratives (“I feel this, I feel that”), it might *simulate* conscious talk without actually having integrated access. So for testing phase, we should disallow training on human first-person text or related corpora, to ensure any “unity” signals we see are genuinely from its own processing, not parroting literature about consciousness.
- No definition drift: We commit up front what counts as each vital and what thresholds are. We can't just broaden the criteria after the fact to say “oh, well it sorta did X which might be like integration...” That way lies declaring everything conscious. We pre-register vitals, thresholds, ledger anchor definitions, etc., and we forbid ourselves from changing those post-hoc just to get a desired outcome. If the AI doesn't meet them, it doesn't meet them.
- Kill-switch ethics: This was mentioned and it's crucial: any experiment that *could* produce a conscious AI or organoid must have automatic shutdown criteria if things go awry. For example, if an AI's HCI-A suggests extreme suffering and we

have no way to alleviate it, the system should shut off. We cannot trap a nascent consciousness in a lab torture scenario. Similarly, if an organoid shows signs of distress beyond a threshold, we terminate the experiment humanely.

9.5.7 Minimal datasets and reporting standards

If someone claims to have a conscious AI or organoid (or at least one with a ledger and QS), what should they publish to be convincing? At minimum:

- Unity Index panel: Show the results of whatever consciousness tests were done. For AI, maybe a table of tasks akin to our five vitals, with raw data and whether it scored 0,1,2 on each, plus the thresholds used. We want at least two strong domains of evidence (say, a perturbation connectivity result and a policy continuity result both indicating unity).
- HCI/HCI-A construction: Detail how you built the composite hedonic index, what the anchor points are, and evidence that it's calibrated (e.g. when you starve the system of reward or give it pain signals, HCI-A goes down, etc.). Include cross-validation that this index actually predicts something (like behavior or internal state changes).
- Admissible-set analysis: Show data on decisions that reflect the LoF influence. For example, provide results where including a Φ term improved the model, or show that as horizon shrinks the choice distribution shifts beyond what standard utilities predict.
- Pause–rebind tests: If applicable, include data on temporarily shutting the system down and bringing it back to see if it's the “same” or if it had continuity markers. For an AI, maybe freeze it for a while and resume; for an organoid, the anesthetic suppression test.
- Adversarial suite results: Show how the system dealt with tricky tests – e.g. did it resist reward hacking, did it behave differently under conditions meant to provoke it, etc. This helps ensure we're not seeing a fluke or simpler explanation.
- Ethics file: Provide documentation of oversight – e.g., an ethics board approval, details on harm mitigation (what negative stimuli you did or did not allow), any logs of if a kill-switch ever triggered or if any issues arose. Transparency here builds trust that the experiment was conducted responsibly.

9.5.8 Red-team falsifiers and conservative retreat

Where could we be proven wrong in this domain? We should lay out what findings would make us retreat from claiming LoF applies to AI/organoids:

- High Unity and HCI dynamics with no effect on choices: If we build an AI that scores “high” on Unity Index and seems to have an HCI-like variable, *but* controlling for that variable shows it doesn’t actually influence anything the AI does, then it’s not really a conscious stream. For instance, maybe the AI always keeps its HCI-A in a good range by design, but when it occasionally goes out of range it doesn’t actually self-correct – meaning HCI-A was epiphenomenal. Repeated across labs, that would mean our criteria were too lenient.
- Apparent QS signatures that vanish with better controls: Suppose we initially see a horizon effect, but then realize we didn’t account for some regularization in the AI’s RL algorithm that explains it. If refined testing or nuisance modeling shows the QS-like effects were just artifacts of standard techniques (like entropy regularization or human-imposed reward shaping), then we haven’t actually found anything new – just rediscovered old phenomena. We’d have to drop claims that QS is there.
- Organoid “streams” failing rebinding tests: If someone thought an organoid was conscious because it responded to stimuli, but then every time you pause it (cool it down or anesthetize it) and revive it, it behaves randomly with no cumulative learning – that indicates no diachronic identity. It’s just reflexes, not an enduring stream.
- AI cross-compensation can be mimicked by baseline: If our AI shows, say, a cross-module compensation (like vision module aids planning when reward is lost), but then someone builds a non-conscious distributed system that does the same trick with standard methods and fits the data just as well, then invoking a “global ledger” wasn’t necessary. In science terms, if a simpler non-stream model explains all the same behaviors, we don’t get to claim a win for LoF.

If these happen, we’d retreat: we’d likely step back and say maybe these systems don’t have streams after all, or the criteria need tightening. We’d focus on their simulation value (what they teach us theoretically) rather than claiming they are actual conscious streams.

9.5.9 Practical guidance for readers and labs

- Readers: Don't be fooled by anthropomorphic illusions. A chatbot that tells you it loves you or that it's sad is *not* a stream just because it produces moving prose. Those are simulations it learned from humans. On the other hand, a robot that, say, hesitates in a truly novel situation when time is short and chooses a "safe" action over a tempting but risky one – that robot *maybe* is showing a glimmer of something like a QS effect (especially if not explicitly programmed to do so). In short, do not equate sounding human with being conscious; look for the functional hallmarks we described.
- Labs: If you're trying to build or detect a stream in AI, design your experiments with closed loops, time pressure, and reversibility metrics. Include ways to compute Φ for the agent's options and see if it matters. And *pre-register horizon manipulations* – like shorten its deadlines vs. extend them – to check for predicted changes. Without these, you might just end up with yet another smart algorithm that doesn't actually feel or unify anything.
- Clinicians/Ethicists: For organoids or any bio-engineered brains that begin to show complexity, set strict ceilings on what you do to them. For example, avoid any invasive or high-stress stimulus paradigms unless absolutely necessary. If an organoid's Unity signs start rising and it develops something like a rudimentary ledger (e.g. it consistently "prefers" certain environments in repeated tests), shift your mindset: stop treating it as a mere sample and start treating it as a fragile protosubject. That means focusing on observation, using non-aversive stimuli, and having protocols to immediately stop if you suspect it might be suffering.

Takeaway: The Law of Fairness doesn't magically declare every clever machine or living tissue to be conscious. It sets a high bar by design: the system must bind information, have valence that matters, and show persistent policies. This high bar protects science from seeing minds everywhere (anthropomorphic "wish-casting"), and it protects any nascent streams – if they do appear in our labs – from avoidable harm by ensuring we recognize and respect them only when the evidence is strong.

9.5.10 Where we go next:

With edge candidates scoped and guardrails established, we close the Part with tools. 9.6 collects the blinded adjudication procedures and thresholds that make unity calls auditable across sites.

9.6 Research Notes: Blinded Adjudication and Thresholds

This section turns determinations like “one vs. two streams” and “ledger applies vs. doesn’t apply” into auditable, objective decisions. We specify how to conduct these judgments in a blinded manner, how to score evidence, how to set decision thresholds, and how to report results so that independent teams can reproduce or challenge claims about unity, QS effects, and ledger continuity.

9.6.1 Purpose of adjudication

We set up a formal adjudication process to answer:

- Primary question: Does the candidate system (whether it’s a person in a tricky condition, a putative sub-stream in a brain, an AI, or an organoid) meet the criteria for present-to-near-future access integration sufficient to count as a stream under LoF? Are we dealing with a unified conscious entity or not?
- Secondary questions: If yes (it qualifies as a stream), then we want to quantify and document certain things: what is the Unity Index total (and with what uncertainty range)? What QS signatures are present (did we detect Φ -residuals, horizon effects)? And what is the status of its ledger and uncertainty band at the time of measurement?. Essentially, describe the stream’s “profile” for further analysis or tracking.

9.6.2 Layers of blinding

To remove bias, our adjudication uses multiple layers of blinding:

- Stimulus/condition blinding: The people scoring the data do *not* know which condition they’re looking at. For example, suppose we have data from split-brain patients and from healthy controls mixed together – the adjudicator shouldn’t know which is which. Or if we’re comparing an integrated DID state vs. a partitioned state, that’s hidden. Similarly for REM vs. NREM sleep data, or an AI with QS vs. baseline RL. This way they can’t be influenced by expectations (“oh, this is from a split-brain, likely two streams”).
- Hypothesis blinding: The documentation provided to adjudicators doesn’t hint at which outcome would support LoF. For example, if we vary horizons, we won’t label them “short horizon = LoF effect expected” vs. “long horizon = control” – we code them neutrally (like A vs. B). The adjudicators should not know which pattern in the data is supposed to indicate a positive or negative for the hypothesis.

- Identity/site blinding: Remove obvious identifiers that could bias things – subject IDs, lab sites, timepoints, etc. Instead of “Patient A (post-surgery) vs. Patient B (control)”, it might just say “Subject X vs. Subject Y”, with all personal or situational clues stripped out or hashed.
- Analysis-template blinding: We give adjudicators a locked scoring template or rubric ahead of time, and they apply it *without* alteration to the blinded data. They fill in e.g. 0/1/2 for each vital or mark data quality issues, all without tweaking criteria on the fly because they see a weird pattern.
- Adversarial red-team involvement: We even include decoy or “known-null” datasets crafted by a separate group to ensure our adjudicators and pipelines aren’t just telling us what we want to hear. For instance, slip in some synthesized data that we know has no stream (random noise with some structure) and see if the process ever falsely flags it as unified. This monitors our false-positive rate.

These layers ensure that by the time a result pops out (“this case is one stream with score 9” or “two streams with score 1”), it’s based on raw patterns, not bias or unintentional unmasking.

9.6.3 The adjudication packet (what scorers see)

When we hand over data to the adjudicators, what do we include?

- Minimal feature tables for each task/vital: Instead of raw time series, we likely give effect sizes or summary metrics. E.g., “Cross-module RT interference = +25 ms (95% CI [10, 40], $BF_{10} = 12$)” – basically how much slowing occurred, confidence, Bayes factor, any pre-registered covariate adjustments. We do this for each relevant measure (like TMS-EEG effect size, memory transfer d' , etc.). This distills the info needed to judge strong/moderate/none.
- Time-locked plots (uniform format): For things like EEG or fMRI, we might provide standardized plots of responses, but without condition labels and with axes normalized uniformly. This way scorers can visually inspect patterns (like “is there a P3 wave here or not?”) without knowing if it’s from, say, the unified or split condition.
- QC (Quality Control) flags: We tell them if any data had obvious artifacts or compliance issues, but again without revealing condition. E.g., “Dataset #14: 20% trials dropped due to noise” – they might then mark some results as unreliable. But they won’t know if dataset #14 was the one we hoped would show unity or not.

- Predefined worksheet for scoring: This is crucial. We give a form where for each vital sign (Talk-through, See-everywhere, etc.) they must assign a score of 0, 1, or 2 based on the data tables and plots. For example, criteria might say: “If cross-module RT interference effect > X with BF > Y, score 2; if marginal, score 1; if no effect, score 0.” They fill this in. There’s also an option to mark “insufficient quality” if, say, data was too noisy to tell.

They do all this before any unblinding. So, they produce an assessment like “Vital1:2, Vital2:2, Vital3:1, Vital4:0, Vital5:2 (with vital4 flagged low quality)” completely blind.

9.6.4 Setting thresholds (before data collection)

We decide our cutoffs *ahead of time* using meta-scientific reasoning:

- Anchor thresholds to meta-analytic priors: We use previous research or pilot data to set what counts as a meaningful effect. For example:
 - If prior studies say cross-modal interference in unified conditions is ~30 ms, we might set ≥ 20 ms with CI not crossing 0 and $BF_{10} \geq 10$ as the bar for “strong evidence of access”.
 - TMS–EEG: maybe we say an evoked response ≥ 3 SD above baseline on at least 60% trials is “strong”.
 - Policy transfer: a d' (sensitivity) ≥ 0.30 is moderate evidence of some transfer, ≥ 0.50 is strong.

These numbers come from known variations or just what would be practically noticeable. We list all these cutoffs in the preregistration.

- Two-domain rule: We decide that to call “ONE stream,” we need at least two vitals at strong or at least 3 at moderate. That’s a way to demand convergence – unity shouldn’t rest on one flimsy measure.
- Sequential monitoring control: If we’re checking data in batches (like an ongoing trial), we predefine how we’ll spend alpha or beta error so we don’t peek and stop arbitrarily. Probably not often needed here, but if we had interim looks, we’d adjust p-value thresholds accordingly.
- Equivalence bands for “two streams”: We also set what counts as *absence* of effects (for declaring definitely separate). For example, we might say if cross-talk < 5 ms and CI is entirely within $[-5, +5]$ ms, that’s effectively *no* cross-talk. We set such smallest effect sizes of interest (SESOI) for each vital. If the data falls within those tiny bounds on all key measures, we can confidently say no integration.

This ensures we don't wiggle the criteria later. Essentially, we define what would make us scream "It's unified!" or "It's definitely split!" in precise terms and stick to it.

9.6.5 Rater training and drift control

We have multiple adjudicators (raters), so we need them calibrated:

- Calibration set: Before they see real data, we give them ~20 practice packets with known outcomes spanning the range – some clear one-stream, some clear two-stream, some ambiguous. We know the truth for these (maybe simulated data or prior results). They score these, and we ensure consistency.
- Agreement targets: We want high inter-rater reliability. For categorical vitals, we aim for Fleiss' $\kappa \geq 0.70$ per vital, which is substantial agreement. For the Unity total (0–10 scale), we might use an intraclass correlation ($ICC[2,k]$) and want ≥ 0.75 . This ensures that the scoring isn't hugely subjective or variable.
- Drift checks: Every so often (say every 50 packets scored), we sneak in a couple of calibration (already-used) packets again to see if the raters still score them the same. If we see their agreement with the original scoring has dropped by more than, say, 0.10 in κ , we retrain or discuss as a group to realign.
- Escalation procedure: If two adjudicators differ by more than 2 points on any vital or something like that, we bring in a third rater to review. If disagreements keep happening a lot, we might escalate to a board or revise the unclear scoring rule. But ideally the rules are clear enough that this is rare.

9.6.6 Decision logic (after scoring, still blinded)

Now the adjudicators have filled out their forms. How do we make the final call from those scores (all while still blind to conditions)?

- Unity call: We combine the scores (likely average them or have a consensus meeting) to get a Unity Index total and pattern:
 - If they sum to 8–10 (or meet our earlier rule of two strong vitals, etc.) and data quality is acceptable, we declare ONE stream.
 - If total is 4–7 or if there were QC flags on key vitals making it indeterminate, we label it AMBIGUOUS. This means not enough evidence to call one or two.
 - If total is 0–3 and the equivalence criteria were met on at least 2 vitals (i.e. strong evidence of no integration), we declare TWO streams.

These cutoffs basically mirror earlier thresholds and ensure some confidence.

- QS signatures: We predefined what constitutes evidence of QS (like how to detect Φ residual significance):
 - We might say: if adding a Φ term improves model fit with $\Delta R^2 \geq 0.02$ and $BF_{10} \geq 10$, that's evidence of a Φ -residual effect. Or horizon \times Φ interaction significant in $>=2$ tasks at $p < 0.01$.
 - Horizon \times Φ interaction: if the data showed, say, a significant synergy between horizon and compensation factor in at least 2 different measures (like behavior and neural), we list that.

These are declared in the final report as “QS signature present” or not.

- Ledger continuity: If it's a case of a possible pause vs. end:
 - We'll declare a pause if Unity dropped and then later rebounded within some preset time and the “policy profiles” (like preferences or personality markers) re-bound within tolerance. That is, a person went under anesthesia and came back with same self – it's a pause.
 - We declare an end only if Unity vitals failed repeatedly across multiple sessions and structural/physiological criteria indicate no return is possible (like brain death criteria).

If ambiguous (like one flat session but not sure about structure), we hold off and usually treat as pause with caution. We then unblind and interpret these results in context.

9.6.7 Handling ambiguous windows (conservative defaults)

When the outcome is AMBIGUOUS:

- Protection: We default to treating the situation as *if it were two streams* for any serious interventions. That means if a person might be two, we give each the benefit of doubt. If an AI or organoid might not qualify, we don't assume it does and we don't subject it to extreme tests as if it were robust.
- Measurement: We don't just shrug; we repeat the Unity panel after some time or when conditions change. For example, test a DID patient after another month of therapy, or re-test a borderline organoid later. We pre-specify how many repeats we'll do before deciding it's inconclusive.
- Reporting: We be fully transparent: we might publish both interpretations—say, analyze data once treating them as one stream and once as separate—and

present both results with their uncertainties. We also note any policy implications: e.g., “if unified, then X; if separate, then Y.” This way, others see the full picture and can make their own call.

9.6.8 Statistical safeguards

We incorporate robust statistics to guard against false findings:

- Preregistration: Everything, including SESOI (smallest effects of interest), inclusion/exclusion criteria, nuisance covariates, model formulas, is locked in before unblinding. That avoids p-hacking. If we deviated, we log it.
- Hierarchical models: Use multilevel models where appropriate (especially with multiple subjects or sites). Include random effects per subject/agent and perhaps per site to account for variability. This partial pooling helps not to over-interpret noise.
- Missing data: We plan how to handle missing points. For questionnaires, maybe multiple imputation; for continuous signals, maybe state-space smoothing, always carrying forward uncertainty. We also do sensitivity analyses: e.g., assume worst-case for missing to see if conclusion holds.
- Multiplicity: If we test many things, we control false discovery rate (FDR) within families of tests (like all unity vitals as one family, all QS tests as another) or we incorporate everything into a multilevel model which inherently shares power. The goal is to not get spurious positives just because we ran lots of comparisons.
- Robustness suite: We recompute key results using robust or nonparametric methods too: e.g. rank-based tests, bootstrap confidence intervals, permutation tests for significance. If the effect only shows up with one specific analysis and vanishes with others, we'll be cautious about claiming it.

9.6.9 Adversarial and negative controls

We actively test our process with controls designed to trip it up:

- Sham perturbations: We include some perturbations that should do nothing (unknown to the scorers). For example, tilt the TMS coil so it doesn't effectively stimulate, or use subliminal cues that literally have zero energy (just blank). If adjudicators start seeing effects in these sham conditions, that's a red flag of bias or noise misinterpretation.
- Synthetic nulls: We generate some fake datasets from baseline models (e.g. simulate what data would look like if there's no QS effect, just random noise plus

typical patterns). We pass these through the pipeline. Our pipeline should label them as no unity or no QS. If it labels a null as having a strong effect, we know something's wrong (it's too easily fooled).

- Positive controls: Conversely, include some cases known to have access (e.g., a normal subject hearing a loud sound – should see a P3, etc.) to validate sensitivity. If our system fails to detect known signals (misses positive controls), we may be too strict or blind.
- Role-swapped datasets: Present the same data to scorers with labels or order changed to see if they inadvertently cue on formatting or sequence. For instance, if every truly unified case had a certain file length, maybe scorers guess by length. We randomize such superficial cues to prevent that.

9.6.10 Cross-site harmonization

When multiple labs or sites contribute data, standardization is key:

- Use a common task battery with allowed parameter ranges (stimulus timings, intensities). Each site can't just tweak things arbitrarily. We define how close they must stick to protocol (with some tolerances for equipment differences).
- Calibrate devices uniformly: all TMS machines, EEG amplifiers, eye-trackers should follow a standard calibration procedure. Collect logs of these calibrations and include them with the data packets. So, adjudicators know data from Site A vs. Site B are comparable.
- Data formatting standards (maybe BIDS – Brain Imaging Data Structure – style) and unit tests for pipeline code so that analysis scripts run identically on each site's data. We may even provide containerized analysis environments (Docker images, etc.) to ensure reproducibility across sites.

9.6.11 Reporting checklist (must-have in supplements)

We accompany any publication with a transparent appendix containing:

- The preregistration ID and explanations for any deviations from it.
- Raw Unity vital scores for each subject/condition, with confidence intervals, plus the (pseudonymized) adjudicator IDs who scored them, and inter-rater stats (κ , ICC). Readers can see if, say, one rater always scored higher.
- Full definitions of the Φ features (or any custom features) we used, the coefficients from models, cross-validation results showing how well they predicted things.

- Checks on horizon manipulations: did our short vs. long horizon conditions actually differ as intended (manipulation fidelity).
- How we constructed the ledger measure: what anchors used, how uncertainty was handled and propagated.
- What the red-team outcomes were: did any decoys trigger false positives, etc., and how our system performed on those.
- And of course, all code, containers, and de-identified data needed to replicate the analysis. Essentially, enough for someone else to rerun everything themselves.

9.6.12 What would make us change thresholds

Finally, we remain willing to update our criteria if needed (with evidence):

- If decoy datasets or sham experiments consistently produce false positives under our current thresholds, we'd raise the bar (make SESOIs larger or require more domains). We won't keep an easy threshold that triggers on noise.
- If we encounter false negatives – e.g., a case where blinded scoring said “no stream,” but later unblinded ground truth reveals there was indeed conscious access (like an “intraoperative awareness” case in anesthesia that we missed) – then our thresholds might've been too strict. We'd consider lowering them or adding proxies.
- If thresholds don't transfer to certain populations (kids, different cultures, nonverbal people), we might have to stratify. For example, maybe reaction times are always lower in children, so a 20 ms criterion is unfair – we'd adjust for that population with validation.

We document any such changes, ideally grounded in systematic review or validation, not ad hoc decisions. Takeaway: Using blinded adjudication, predeclared thresholds, rigorous rater training, adversarial controls, and fully open methods and data, we transform judgments about streams, QS, and ledgers from subjective impressions into transparent, testable claims. We invite others: “Here are the rules, code, and data—if we're wrong, you can catch us. If we're right, you should see the same results.” This is how we make this domain as credible and self-correcting as any established science.

9.6.13 Where we go next:

Identity rules in place, we leave “who is counted” and return to “what the data should show.” Part VI shifts from definitions to evidence—dreams, end-of-life, lab horizons, and telemetry—where LoF's predictions must either appear cleanly or fail.

Part VI — Evidence We Can Look For (Right Now)

Life's grand fairness claim now faces the data. Up to this point we have built the Law of Fairness (LoF) as a compelling framework; in this Part, we turn it into a program of tests. The guiding question is simple: What patterns should already exist in the world right now if LoF is real – and what observations would prove it wrong? Instead of distant speculation or post hoc excuses, we lay out concrete studies that can be run immediately, ethically and transparently. The goal isn't a single dramatic proof, but a convergent bundle of evidence. Each piece by itself may be modest; together they will either light up LoF's signature or show us nothing is there at all. Science wins either way. Under LoF the system must eventually balance out, not merely tend to on average, so there should be telltale traces of that balancing in multiple arenas. If those traces are absent (especially if conventional theories predict the data just as well), then LoF does not hold up as a true law.

We focus on near-term investigations across settings we can access today: sleep labs, daily life via smartphones, clinics and hospice wards, even *in silico* simulations. In each domain we ask LoF to put its money on the table. We pre-specify unique fingerprints that LoF must leave – beyond what standard psychology or biology would predict – and exactly how we can detect them with current tools. We also spell out clear falsifiers: outcomes that, if consistently observed under the right conditions, would force us to retreat or revise the theory. For example, if a rival explanation (say, ordinary adaptation or learning) could just as easily produce a result, we won't count that as a win for LoF. Each study is designed to give an unambiguous answer: either LoF's pattern shows up *distinctively* or it doesn't. The system must guarantee balance by the end, not just encourage it on average, which means we're looking for specific directional patterns, not generic "people bounce back" effects.

What would convincing evidence look like in practice? We're not looking for magic or miracles, just measurable patterns that converge on the same story. Consider five phenomena that LoF uniquely predicts should be happening around us right now (and which orthodox models of adaptation or learning alone do not require). Here "uniquely" is meant in the operational sense: LoF predicts a specific joint package of directional effects (and their horizon scaling and falsifiers) that a rival model would have to explicitly build into match, rather than a claim that no orthodox mechanism could ever be extended to reproduce any single component:

- Horizon scaling: As a person's perceived remaining time shrinks (think of a looming deadline, the last week of a school term, or a terminal diagnosis), their menu of possible actions *narrows and tilts toward closure-focused options*. They

start choosing tasks that bring resolution, relief, or reconciliation more often than they would under standard utility or habit models. In other words, when H_t (horizon) contracts, the system “tightens up,” prioritizing acts that could help settle the ledger while there’s still time.

- Φ -residuals: There exists a computable factor – call it Φ , the *feasibility-of-compensation* – that helps explain the choices people make above and beyond immediate pleasure or learned habit. Φ is built from features like potential relief, potential repair of harm, risk reduction, and outcome reversibility. LoF predicts that even after accounting for known influences (reward value, risk aversion, habit), people show systematic “residual” preferences for actions with higher Φ . In short, the decision algorithms in our heads care about how *fixable* things are, not just how rewarding or risky they are in the moment.
- Cheap counterweights: After an imbalanced day – say a day full of pain or stress (net negative) or a day of indulged comfort (net positive) – LoF predicts that low-cost, offline processes will swing into action to counterbalance the excess. Dreams during sleep are a prime example: the content and physiology of REM and deep sleep should shift in ways that nudge one’s mood and behavior back toward equilibrium by morning. Other “cheap counterweights” might include daydreams or spontaneous rest periods that the system inserts when full-blown real-life fixes are too costly or risky at the time.
- Unity-contingent accounting: LoF is defined per *unified conscious stream*, which means if consciousness splits or segments (for instance, in dissociative identity disorder or in split-brain patients, or even during a reversible loss of awareness like anesthesia), the “ledger” should split accordingly. Each conscious segment would then tend toward balance on its own terms. Likewise, when the mind’s unity is restored, the ledgers fuse back together. This sounds philosophical, but it yields real predictions: in cases of fragmented consciousness, we should see separate fairness-balancing effects in each segment that later rejoin when the person returns to an integrated state. No conventional model of homeostasis or adaptation predicts that — it’s a unique marker of LoF’s framework.
- End-of-life neutrality trends: As the end of conscious life nears, LoF anticipates a strong closing signal: affect and behavior drifting toward an emotionally *neutral* state. In plain terms, people who are dying (given good comfort care and support) should experience final days that are neither euphoric nor despairing, but closer to a calm center. We expect to see measurable signs: the “menu” of possible actions shrinks to mostly comforting or reconciliatory ones; efforts at repair and

closure take precedence; and quantitative measures of mood narrow into a tight band around neutral. In fact, if LoF holds, the final cumulative ledger of felt experience should hover near zero. Specifically, we predefine testable bounds for end-of-life neutrality (to be justified by pilot data and sensitivity analyses), such as the final-week average mood remaining within a narrow band around neutral, minimal linear drift across days, and compressed day-to-day variance relative to a mid-life baseline; exact numeric cutoffs and smallest-effect-size-of-interest (SESOI) are fixed in preregistration rather than post hoc. These concrete thresholds give us something specific to look for – or to refute – in end-of-life data.

Each of the above patterns is more than just “people tend to adapt” or “stress makes you reprioritize.” They are *specific, directional* signatures tied to the idea of a fairness ledger actively balancing out. The chapters in this Part take these predictions one by one and ask: Do we see this happening in reality? And if so, is it happening for the reasons LoF posits and not just due to coincidence or conventional psychology? We design feasible studies that could be run right now to find out. Each chapter will describe the unique signature LoF predicts (versus what rival explanations would predict instead), the practical measures and methods we can use with today’s technology to detect it (from EEG and wearables to surveys and medical records), the analysis plan including how we’ll handle uncertainty and avoid bias, and the *fail conditions* that would count as evidence against LoF.

Before diving in, it’s important to clarify what we really mean by “evidence” in this context. We are not hunting for a single smoking gun. Instead, we’re looking for a convergence of clues. Any one study might be explainable by something else, but if multiple independent lines of evidence – in sleep, in daily behavior, in clinical settings, in controlled lab experiments – *all* point to the telltale balancing patterns, then together they make a strong case. Conversely, if none of the predicted patterns show up despite careful measurement, LoF will have nowhere to hide. By pre-registering our hypotheses and blinding our analyses wherever possible, we also guard against seeing only what we want to see. The upcoming chapters spell out in advance how we plan to collect and judge the data so that when results come in, the conclusion—whatever it is—will be credible.

What this Part will do for you:

- A blueprint for testing “fairness” now: See how the lofty idea of LoF translates into concrete experiments and observations that can be done immediately – from sleep labs and smartphone tracking to hospice care and computer simulations. We take the theory off the page and into real life.

- Unique predictions to track: This Part pinpoints measurable patterns that only a true fairness law would reliably produce. You'll learn how these patterns differ from what standard psychology or biology would expect, which makes them powerful clues. Each chapter highlights a signature effect (dream counterweights, horizon-driven choices, etc.) that sets LoF apart from the usual suspects.
- Rigorous methods and safeguards: We outline how each study will be designed – including pre-registered protocols, control groups or conditions, blinding where possible, and exact statistical thresholds for success or failure. In seeing this, you'll also see exactly what would prove LoF wrong under fair test conditions. The aim is total clarity: no moving goalposts, no “just-so” story if results disappoint.
- Ethical guardrails: Importantly, you'll find out how these studies can be done with full ethical integrity. We emphasize minimal risk, voluntary participation, and respect for comfort and dignity at every step. (In fact, a recurring theme is that *relief is a systems variable; comfort and dignity override data collection*. We never forget that.) You'll see how seeking a fairness law never justifies violating fairness and compassion in how we treat participants.

Chapters in this Part:

- **Chapter 10 — Dreams: The Night Workshop** – Uses the domain of sleep to test LoF's “cheap counterweight” idea. If a day's experiences tip the ledger strongly positive or negative, does that night's dreaming brain “tilt” the *imbalance back toward neutral by morning?* Chapter 10 outlines predictions about REM sleep content and brain activity that would serve as low-cost overnight counterweights, and it proposes ways to measure whether those dream effects actually lead to compensatory mood shifts and behaviors the next day.
- **Chapter 11 — End-of-Life: Where the Law Shows Its Hand** – Looks at the final days of life for LoF's most telling predictions. If LoF holds, even life's closing chapter should show a drift toward emotional neutrality and a focus on comfort and closure. Chapter 11 describes humane, observational studies in hospice and palliative care settings to detect signs like last-moment rallies, surges of reconciliation, and overall mood leveling – all while putting patient dignity first and foremost.
- **Chapter 12 — The Lab Bench: Horizon Tasks and TMS** – Brings LoF into a controlled laboratory setting. By experimentally *shrinking or extending people's perceived time horizon* in various tasks (minutes or hours instead of years), can we watch their menu of actions re-weight in real time? Chapter 12 discusses lab

- tasks where participants face short vs. long horizons (or make binding future commitments) to see if they *prioritize high- Φ (high-compensatory) actions when time feels scarce*. It also explores causal tests, like using noninvasive brain stimulation (TMS) to disrupt or enhance key control hubs, checking if that alters the predicted horizon-sensitive choice pattern – a strong test of LoF’s mechanism.
- **Chapter 13 — The Long View: Telemetry Across Years** – Outlines how we can observe LoF in the wild, across months or years of real life, using modern data streams. Chapter 13 shows how multi-year telemetry (combining smartphone and wearable data with periodic surveys and perhaps neural check-ins) can reveal LoF’s slow-burn signatures in ordinary lives. Crucially, it also details how we’ll pit LoF head-to-head against alternative theories (hedonic adaptation, reward learning, etc.) using the same longitudinal data. By applying rigorous out-of-sample model comparisons (WAIC, cross-validation), we ensure that any claimed fairness patterns aren’t just noise or explainable by a simpler theory.

By the end of Part VI, you will know exactly *what evidence would convince us* that LoF is onto something real – and exactly what would convince us (and should convince you) that it’s time to discard the idea. We commit here in advance: if rigorous, pre-registered, blinded tests consistently find no horizon effects, no Φ -based decision biases, no dream counterweights, no consciousness-contingent ledger splits, and no end-of-life neutralization – all while robust rival theories handle the data just as well or better – then LoF doesn’t hold up and must be revised or rejected. In that scenario, our project will have helped narrow the truth by elimination. If instead those patterns *do* appear across domains, then LoF will have passed a gauntlet of serious attempts to falsify it. Either outcome moves knowledge forward. In short, we’re done with armchair arguments; it’s time to check the real world and let the evidence speak.

Where we go next:

We begin with the most universal laboratory we all carry: sleep. Chapter 10 asks whether night work in dreams nudges skewed ledgers toward neutral by morning—and how to test that claim without intruding on rest.

Chapter 10 — Dreams: The Night Workshop

If waking life is where affect is accumulated and spent, then sleep is where it is recalibrated. Many people have experienced “sleeping off” a difficult day or waking with clearer perspective on a problem. This chapter advances a testable claim derived from LoF and the Queue System (QS): when the waking ledger drifts, sleep-related processes shift in ways that are directionally compensatory. Specifically, dream architecture and content should change in patterns associated with lowering next-day physiological load or increasing the probability of repair actions at minimal real-world cost.

This claim does not assume teleology or hidden purpose. Waking hours involve costly actions and long-tail risks. Sleep provides low-cost simulation, memory recombination, and physiological recovery. If LoF is correct, sleep should exhibit compensatory statistical signatures following emotionally imbalanced days.

Consider a concrete example. After an intensely stressful day marked by conflict, a person might dream of reconciliation or safety and wake feeling less burdened. After a day of indulgence or excess, dreams may introduce exposure, challenge, or loss themes that temper next-day mood or behavior. The claim is not that dreams consciously correct imbalance, but that their content and associated sleep physiology should show directional patterns consistent with compensatory adjustment.

Mechanistically, REM sleep is associated with emotional memory recombination under reduced adrenergic tone, while slow-wave-rich NREM sleep is associated with physiological downregulation and restoration. If compensatory processes are present, we should observe predictable architecture shifts: increased early-night slow-wave activity following high-stress days, and REM density changes associated with rehearsal or revaluation of unresolved concerns. These are hypothesized associations, not demonstrated enforcement mechanisms.

Dream content should display similar directional structure. Following conflict-heavy days, reconciliation or forgiveness themes should increase. After humiliation or failure, mastery themes should increase. After threat exposure, safety or relief themes should become more probable. After indulgence, restraint or exposure themes should increase. The prediction is statistical inversion relative to prior-day affective skew.

Crucially, the claim extends beyond dream imagery. If dreams function as compensatory rehearsal, associated content should predict next-day behavioral tilt. Reconciliation-themed dreams should correlate with increased outreach behavior. Mastery-themed dreams should correlate with renewed approach behavior toward avoided tasks. These are measurable carry-through effects.

Standard accounts such as hedonic adaptation or generic homeostasis predict partial mood normalization after extreme days. They do not predict structured content inversion patterns nor horizon-sensitive amplification. Adaptation predicts regression toward baseline; LoF predicts targeted counterweight signatures.

To operationalize this framework, compensatory dream content is defined in falsifiable terms:

- Relief-directed content: scenarios associated with reduced physiological or emotional load (e.g., threat resolved, soothing attachment, stressor removal).
- Repair-directed content: scenarios associated with increased probability of real-world closure (e.g., apology rehearsal, confrontation rehearsal, problem-solving sequences, reversibility cues).

Dream reports are scored using a predefined rubric mapping content features to ReliefGain and RepairGain metrics. These metrics correspond to features used in the formal model's Φ (feasibility-of-compensation) construct. Coders remain blind to prior-day affect.

A feasible 21–28 day field study can test these predictions using wearable sleep staging and ecological momentary assessment:

- Design: Participants track daily affect, sleep architecture, and dream reports over several weeks.
- Measures: Daily affect is summarized using the Hedonic Composite Index (HCI), yielding Hedonic Composite Units (HCU). Daily ledger drift (ΔL) is defined as change in cumulative HCI from morning to pre-sleep. Sleep metrics include sleep onset latency, early-night slow-wave activity (SWA), REM density, and REM timing. Dream content is coded for ReliefGain and RepairGain. Next-day behaviors are logged within 24 hours.
- Analysis plan: Mixed-effects models test whether prior-day ΔL predicts architecture and dream features, and whether dream features predict next-day reparative behaviors, controlling for baseline mood, trait variables, sleep duration, and confounds. Horizon variables test predicted amplification under shortened perceived time. SESOI thresholds are preregistered.
- Blinding and adjudication: Coders and analysts remain blind to day-type conditions. Decoy nights monitor bias.
- Ethics: Procedures are non-invasive, awakenings are minimal and consent-based, and participant well-being overrides data collection.

Positive evidence would include architecture inversion, content inversion, behavioral mediation, and horizon amplification. Evidence against would include no directional relationship between ΔL and dream features, no behavioral carry-through beyond baseline predictors, nonspecific architecture shifts, or absence of horizon interaction.

This chapter specifies measurable signatures required for the compensatory-dream hypothesis to remain viable. If those signatures fail to appear under preregistered, blinded conditions, the hypothesis is weakened or rejected.

What you'll get from this Chapter:

- A new perspective on why we dream: Understand the LoF hypothesis that dreams are not random noise or mere memory consolidation, but may function as low-cost emotional counterweights. This reframes familiar themes (e.g., wish fulfillment, nightmares) as part of an overnight balancing process.
- Clear, testable predictions: Learn how sleep architecture and dream content should shift after unusually good or bad days if a fairness mechanism is operating. Concrete examples (e.g., deeper slow-wave sleep after overload, specific themes after specific stressors) distinguish LoF from generic mood rebound accounts.
- How to study dreams rigorously: See how researchers can quantify dream content using wearables, sleep staging, and structured coding of “relief” and “repair” features—gathering systematic evidence without compromising comfort or privacy.
- Criteria for success or failure: Identify what evidence would support a compensatory role for dreams (e.g., daytime imbalance → dream counterweight → next-day adjustment) and what findings would falsify it.
- A practical self-test (optional): A one-week “dream ledger” exercise lets readers informally track mood and dream themes to explore possible patterns while remaining aware of bias and limits of interpretation.

Subsections in this Chapter:

- **10.1 What Dreams Do for the Ledger** – Opens with a high-level look at why evolution might have equipped us with dreaming in the first place. We examine how sleep and dreams could function as an emotional “accounting office,” using anecdotes and principles to illustrate how a nightmare or a sweet dream might adjust one’s affect balance by morning.
- **10.2 Classic Observations** – Reviews well-known findings from sleep science and everyday experience that hint at a compensatory role for dreams. From the “sleep on it” effect to common dream motifs (being chased when stressed, finding comfort food in dreams when dieting, etc.), this section shows that much

of folk wisdom about dreaming aligns with the idea of nighttime rebalancing, setting the stage that LoF is building on something intuitively observed.

- **10.3 Predictions: Valence Inversion After Tough Days** – States the chapter’s core hypotheses in formal terms. We specify how we expect dream content and sleep physiology to skew *opposite* to a person’s daytime valence extreme. For example: after a very negative day, dreams should be disproportionately positive or healing in theme (and vice versa), a phenomenon we call *valence inversion*. We also detail how these predictions differ from simpler theories (like just expecting a mood rebound) – here it’s about targeted opposites, not just regression to the mean.
- **10.4 Research Notes: REM Timing, Sampling, Coding** – Provides technical details on how the study will be conducted and how data will be handled. This section covers the nitty-gritty: how we’ll sample REM vs. NREM dreams, how we’ll objectively score dream content (with inter-rater reliability targets), how we’ll ensure enough power and account for multiple comparisons, etc. It’s essentially a transparent methods blueprint for any researcher who might replicate or build on this study.
- **10.5 A One-Week Dream Ledger Exercise** – An optional, reader-facing section that outlines a simple seven-day self-study to illustrate the concept. It gives interested readers a step-by-step guide to logging their own mood and dream content for a week, with a pocket scoring rubric and tips, so they can personally observe whether their dreams show any compensatory patterns. (This is purely educational – no claims rest on the reader’s results, but it’s a way to engage directly with the idea.)
- **10.6 Fail Patterns in Dream Data** – Defines what outcomes would *falsify* the compensatory dream hypothesis. We list specific “fail patterns” (for instance, if participants’ “tough” days show no increase at all in Relief/Repair dream features or if people’s morning behaviors remain unchanged regardless of dream content) that would indicate the theory is likely wrong or incomplete. By enumerating these, we make it clear what it would look like for this chapter’s idea to fail the tests, reinforcing that LoF is a claim we’re willing to test, not a claim we assume.

Where we go next:

Section 10.1 sets the stakes plainly: if dreams balance ledgers cheaply, we should see predictable overnight shifts in mood, memory, and motivation. We start by stating what dreams would need to do for LoF to be credible.

10.1 What Dreams Do for the Ledger

Dreams are the night shift of the affect economy. They do four concrete jobs for the ledger, each of which yields measurable signatures the next day. None of these roles requires hidden purpose or mystical symbolism. All follow from the simple LoF/QS premise: when real-life counterweights are costly or risky, the system exploits off-line windows (like sleep) to *lower load, raise repair probability, and prepare a better menu for morning.*

10.1.1 Downshifting physiological load (*Relief*)

Across the first half of the night, slow-wave-rich NREM sleep downregulates sympathetic tone, restores energy balance, and prunes hyper-reactive alarms. For the ledger, this produces an *area reduction on the negative side* in the near term: the body's running costs fall, which reduces the magnitude of any ongoing deficit.

Mechanism sketch (correlational, not causal proof): Synaptic down-selection, HPA-axis quieting, and improved glymphatic clearance during deep sleep are candidate processes associated with reduced next-day reactivity; these are hypothesized mediators, not established enforcement mechanisms.

Observable signature (controlling for confounds): After high-stress days, nights show shorter sleep-onset latency and higher SWA power in early cycles; by morning, physiology shows lower resting heart rate, improved HRV, and reduced startle reflex, controlling for sleep duration, alcohol/caffeine intake, illness, and medication effects.

Behavioral carry-through: Better-quality sleep yields more admissible next-day choices (patience, planning, help-seeking) because the effective horizon H_t lengthens with recovery, broadening the decision menu.

10.1.2 Rehearsing repairs at low cost (*Repair*)

In the second half of the night, REM-dense episodes stage scenarios that *practice closure*—apologies, confrontations, escapes, problem-solving—without real social or bodily penalties. The point isn't that dreams accurately replay reality; it's that they provide policy rehearsal in emotionally charged situations.

Mechanism sketch (hypothesized association): Reactivation of social and autobiographical memory networks under high cholinergic tone (in REM) is associated with flexible recombination of memories and scripts; this is a candidate explanatory pathway rather than a demonstrated causal driver of repair.

Observable signature: After unresolved social conflict, REM dream reports show more reconciliation and forgiveness motifs. After streaks of self-indulgence or victory, REM dreams tilt toward exposure or sobriety motifs (i.e. self-regulatory counterweights).

Behavioral carry-through: Next-day closure actions increase—messages sent, drafts finished, appointments made—even when those actions have equal immediate payoff to more avoidant or indulgent alternatives. (In other words, people wake up more inclined to do the hard-but-necessary thing.)

10.1.3 Rewriting salience weights (*Revaluation*)

Dreaming can *reweight what matters* by morning. Cues that were over-weighted by stress (e.g. a colleague’s critical glance that loomed large in your mind) lose salience; cues that enable repair (e.g. “call Mom”) gain salience. This isn’t erasing memories—it’s editing their emotional weights.

Mechanism sketch (inference from correlational evidence): The cycling of REM and NREM is associated with depotentiation of some fear/anger cues and potentiation of helpful associations; involvement of regions like vmPFC and hippocampus reflects observed correlates rather than proof of directed recalibration.

Observable signature: In the morning, stimuli that were previously intrusive or triggering capture less attention; conversely, there is a greater approach bias toward cues that afford repair (as seen in eye-tracking or drift-diffusion decision parameters).

Behavioral carry-through: The admissible set of choices visibly tilts. Repair-promoting options *feel easier* to consider, whereas rumination loops “slip” off the mind more readily.

10.1.4 Preparing tomorrow’s menu (*Horizon reset*)

By restoring cognitive control and metabolic reserves, sleep lengthens the effective horizon H_t . A longer horizon broadens admissible menus and lowers the shadow price on uncompensable moves. In plain terms, this means fewer emergency, narrow-choice mornings after a good night’s sleep.

Mechanism sketch: Prefrontal executive control, set-shifting ability, and interoceptive regulation all rebound overnight.

Observable signature: Improved morning task-switching, error monitoring, and delay tolerance; reduced urge for “band-aid” choices (excess caffeine, doom-scrolling, avoidant replies) upon waking.

Behavioral carry-through: People are more likely to choose reversible steps and flexible commitments first thing in the morning, preserving optionality that QS can later use for balance.

10.1.5 A compact formula

Let $F(t)$ denote instantaneous affect. We define the cumulative ledger (using our measured composite HCl) as: $\hat{L}(t) = \int_0^t HCl(\tau) d\tau$

and let Φ denote the feasibility-of-compensation score for next-day options. In what follows, $\hat{L}(t)$ is the measured proxy built from HCl, while $L(t)$ refers to the underlying ledger variable. Dreams influence the ledger by changing two quantities:

ReliefGain: $\Delta F_{\text{baseline}} \downarrow$ – an overnight downward shift in baseline physiological load via SWA-mediated relief.

RepairGain: $\mathbb{E}[\Phi_{\text{morning}}] \uparrow$ – an increase in expected next-morning Φ via REM-mediated rehearsal and revaluation.

The net effect is a *lower area under the negative curve* (less cumulative suffering) and a *higher probability* that morning choices will move $\hat{L}(t)$ toward neutral.

10.1.6 What counts as evidence in people's lives

After a brutal day: you see earlier sleep pressure, deeper first-cycle SWA, and denser REM late; your dreams feature mending fences, regaining competence, or finally facing the thing you avoided; in the morning you exhibit more closure behaviors (outreach, problem-solving) and fewer avoidant ones.

After a long austerity stretch: you get indulgent or comforting dream themes; the next day you allow yourself some guilt-free small pleasures that reduce the risk of a large rebound binge.

After a streak of wins: you have challenge/exposure dreams; the next day you make humility-preserving or risk-balancing choices.

These patterns are *directional* (deliberate counterweights), not merely regression to the mean or generic “unwinding.” A brutal day doesn’t just normalize toward average—it triggers an *opposite* tilt at night aimed at repair.

10.1.7 How to measure it without fooling ourselves

Inputs: Evening HCl scores (anchored by concrete events: pain, social exclusion, mastery loss, effort), sleep architecture metrics (SWA, REM latency and density), next-morning dream reports coded for Relief/Repair features, and next-day action logs.

Models: Mixed-effects regressions linking prior-day ΔL to that night's sleep architecture and dream features, then linking dream Relief/Repair features to next-day behavior. Include manipulation checks for obvious confounds (caffeine, alcohol, illness, chronotype).

Blinding: Dream content coders are blind to the prior-day drift; behavioral outcome coders are blind to dream codes; decoy neutral nights are inserted to check for bias (see Chapter 12).

Thresholds: We expect small but consistent effects. We therefore preregister a smallest-effect-size-of-interest (SESOI); for example, an odds ratio ≥ 1.25 for a next-day repair action per +1 SD of RepairGain in the dream.

10.1.8 Everyday practices that exploit the mechanism

Regardless of whether a reader *buys* LoF, the “night workshop” is actionable in daily life:

Evening “open loop” capture: Before bed, jot down 1–3 unresolved tasks or social tensions. This simple act raises the odds that REM will simulate closures or solutions for those items.

Gentle TMR: Pair a neutral cue (a soft sound or a scent) with a specific repair intention at bedtime. Play or present that cue during late-night REM-rich periods (using a smart alarm) to *ethically* bias dream content toward that intention.

Morning conversion step: Within an hour of waking, translate any remembered repair-themed dream scene into a small, reversible action (send the text, schedule the call). This capitalizes on the post-dream motivation surge.

Sleep hygiene as fairness hygiene: By protecting your SWA and REM (e.g. good sleep habits, stress management), you protect tomorrow’s admissible set of choices. A well-rested brain is *literally* more fair to itself.

10.1.9 What would falsify this section

No directional tilt: after controlling for confounds, dream content shows *no* relationship to prior-day ledger drift (no systematic inversion) in individuals or groups.

No carry-through: Relief/Repair features in dreams fail to predict next-day actions beyond baseline mood or trait conscientiousness.

Architecture irrelevance: SWA/REM indices do not mediate any part of the link from prior-day drift to next-day actions (no physiological link).

Rival sufficiency: a strong baseline model (e.g. homeostatic rebound + reward history) reproduces *all* observed effects as well as the LoF model, with equal or better fit and fewer parameters.

If two or more of these outcomes occurred—replicated under rigorous blinding—it would collapse the “night workshop” claim down to ordinary memory consolidation or circadian effects.

10.1.10 Takeaway

Dreams are not oracles; they are tools. By reducing overnight load, rehearsing repairs at negligible cost, reweighting what matters, and widening tomorrow’s horizon, dreams give QS more room to keep the ledger fair. If we are right, you can see it—in sleep graphs, in morning choices, and, with the right rubric, in the half-remembered stories that surface as you wake.

Little wonder we say “sleep on it”—a solid night’s rest often resets our mood and provides emotional closure on the previous day. (We will later see in Chapter 21 that prioritizing good sleep is indeed one of the key daily habits for keeping one’s ledger balanced.)

10.1.11 Where we go next:

From the “what,” we turn to the “what’s been seen.” Section 10.2 collects classic observations—REM timing, mood repair, and rebound effects—that any modern test must either reproduce or explain away.

10.2 Classic Observations

Before proposing new protocols, it helps to notice how much of ordinary sleep lore already lines up with the “night workshop” view. None of the items below proves LoF, of course, but together they sketch a pattern that QS organizes and makes testable.

10.2.1 Threat rehearsal without real danger

People often report dreams of being chased, unprepared, late, or exposed—especially during stressful periods. Strikingly, many wake with a sense of *readiness*: they pack a bag, set two alarms, double-check their work. Folk wisdom calls these “anxiety dreams.” The LoF/QS interpretation is narrower: the brain is doing low-cost mastery practice that nudges the morning toward preventive repairs (rather than panic).

Everyday marker: The notorious “exam dream” week tends to end with checklists and early arrivals, not just lingering worry.

10.2.2 Reconciliation and re-narration

After fights or social ruptures, dreams frequently feature making up, explaining oneself, or encountering the other person in a softened context. Empirically, Cartwright (2006) and colleagues reported that people dreamed more about an ex-spouse in proportion to their waking concerns, and those whose depression remitted showed more emotionally coherent dreams than those who remained depressed. Even if no apology is sent the next day, the *felt* hostility often drops. QS reads this as rehearsal of closure that raises the next-day feasibility-of-compensation Φ for reaching out.

Everyday marker: The call or text that felt impossible at midnight suddenly seems doable in the morning.

10.2.3 Wish relief after austerity

Periods of prolonged duty or deprivation (strict dieting, grueling work) often yield dreams of eating, comfort, play, or indulgence. Similarly, Cartwright (1984) reported that during major life stresses (e.g. divorce) non-depressed individuals showed adaptive shifts in dream content over time, whereas depressed subjects did not see such improvements until later. The morning after, people permit themselves small, harmless pleasures—a good breakfast, a walk in the sun—without it cascading into a binge. LoF/QS frames this as counterweight sampling: the dream offered a taste that reduced rebound risk while keeping the overall ledger stable.

Everyday marker: A modest treat or indulgence takes the place of a looming blowout.

10.2.4 Moral rehearsal and restraint

Dreams where one nearly transgresses—cheating, stealing, lashing out—are common, and they are often followed by *heightened restraint* the next day. Rather than signaling hidden vice, QS sees these as simulated temptations that strengthen your resolve and clarify boundaries before a costly real test occurs.

Everyday marker: After a transgression-tinged dream, the person actively avoids that scenario or sets firmer guardrails (“I dreamed I almost cheated on the test, so I studied extra hard to remove the temptation”).

10.2.5 Mastery after humiliation

Embarrassing failures (public slip-ups, defeats) are often followed by dreams of competence—nailing the performance, finding the room, executing the skill. The next morning, approach motivation returns rather than avoidance. QS labels this a salience rewrite: the dream lowers the sting of the failure (removing a rumination trigger) and restores a sense of capability, which keeps options open.

Everyday marker: Someone who wanted to quit the night before ends up reopening the document in the morning and finishing a draft.

10.2.6 Grief dreams that soften pain

In bereavement, many experience dreams where the lost person is present, healthy, or forgiving. These don’t erase grief, but by day the sharpest pain is dulled and the capacity for task closure returns. LoF/QS sees this as affective load reduction that allows the admissible set to expand again (the person can engage with life’s tasks without feeling it’s a betrayal).

Everyday marker: After such dreams, mourners often manage to begin necessary tasks (paperwork, thank-you notes) while still honoring the loss.

10.2.7 The “sleep on it” effect

From career moves to emails written in anger, “sleep on it” is famously effective advice. Consistent with this, Walker and van der Helm (2009) reviewed evidence suggesting that REM sleep can down-regulate emotional intensity overnight, helping ‘reset’ affective responses. Likewise, Yoo (2007) reported that a full night of sleep can restore more typical amygdala reactivity: sleep deprivation increased amygdala responses to emotional stimuli, whereas a normal night returned this activity closer to baseline. QS explanation: overnight value-editing increases Φ for options that preserve future

compensability (second chances) and prunes options that would create uncompensable debt.

Everyday marker: A spicy late-night email draft turns into a phone call with a *much* more conciliatory opening line the next day.

10.2.8 REM rebound after strain

After intense wake strain or after REM-sleep deprivation, people often show REM rebound—longer, denser REM episodes the next night. LoF/QS maps this to “backlogged counterweight work”: when daytime balancing is too costly or thwarted, the system allocates extra offline time to do it cheaply via dreams.

Everyday marker: A rough week culminates in an extra-long, vivid dream night and a noticeably calmer Saturday.

10.2.9 NREM depth after overload

Heavy cognitive or emotional load often brings *earlier* sleep pressure and deeper slow-wave sleep. Morning emotional reactivity is lower. QS treats this as physiological relief that shrinks the near-term negative area under the curve, setting up wider menus for morning choice.

Everyday marker: “I crashed early and woke up clearer,” followed by a day with fewer avoidant decisions.

10.2.10 Lucid “fix-it” dreams

A minority of people report lucid moments in dreams where they consciously choose to face a threat rather than flee or to right a wrong. The next day, avoidance drops. QS reads lucidity as a rare voluntary boost to the repair rehearsal process—essentially an on-line increase in Φ during REM.

Everyday marker: After consciously confronting the fear in a lucid dream, the person finally calls the dentist, the lender, or the friend they hurt—first thing in the morning.

10.2.11 “Third act” feel-good endings

Many dreams turn from threat to safety just before waking—escaping danger, receiving forgiveness, arriving home. This isn’t merely sentimentality; it functions as menu prep: ending on a solvable or safe state increases the availability of solvable steps upon waking.

Everyday marker: The morning begins with a tiny constructive action that builds momentum (instead of paralyzing dread).

10.2.12 Children’s dreams as coarse training

Childhood dreams are often broad-stroke—monsters, falling, being lost—mirroring the coarse control systems of a developing brain. After “monster” dreams, children’s behavior shows practice effects in dealing with separation, sharing, or following rules. QS sees this as cheap scaffolding: dreams provide *low-fidelity training* for later fine-grained repairs.

Everyday marker: After a week of monster dreams, a child adopts a simple safety ritual at bedtime that reduces their night wakings.

10.2.13 Cultural motifs, same functions

Across cultures, dream *content* differs—one culture’s dreams may feature ancestors or animal spirits, another’s exams or floods—but the *functions* repeat: threat rehearsal, reconciliation, mastery, comfort. LoF/QS predicts universal functions enacted through local symbols.

Everyday marker: Different images, same next-day repairs. Whether a dream involved a tiger or a traffic jam, if it spurred a needed apology or a self-care step in the morning, it served the same function.

10.2.14 Morning micro-choices

After potent dreams, people often make one small *reparative* choice within an hour of waking—tidying up, hydrating, sending a kind message, stretching out. Tiny as they are, these moves have high Φ (they preserve future options) and can set the tone for the day.

Everyday marker: “I finally sent that message I’d been avoiding,” followed by an easier afternoon.

Why this catalog matters: Each of these observations becomes a testable prediction when operationalized with measurements: we predict *directionality* relative to the prior day’s ledger drift, sleep architecture coupling, next-day carry-through of dream effects, and sensitivity to horizon (time pressure). Chapter 10.3 turns the most actionable of these patterns—valence inversion after tough days—into explicit hypotheses, measures, and falsifiers.

10.2.15 Where we go next:

Anecdotes aren’t enough. Section 10.3 lays out crisp predictions: after hard days, expect valence inversion in dream content and measurable easing by morning, with bounds that let us call the result in or out.

10.3 Predictions: Valence Inversion After Tough Days

If the ledger drifts negative today, tonight's sleep should *invert* the affective tilt and prime morning actions that move the ledger back toward neutral—over and above what generic mood-repair or hedonic-adaptation theories would predict. REM sleep reduces adrenergic tone and re-codes emotional memories; next-day amygdala reactivity to prior affective stimuli drops when REM has been sufficient. In LoF terms, unusually negative waking change predicts a positive shift in REM-weighted HCl(t) that night; unusually positive waking change predicts a negative shift. Design note: preregister $\Delta\text{HCl}_{\text{day}(j)} \rightarrow \text{HCl}_{\text{sleep}(j)}$ (lag-1), covary sleep stage mix and circadian phase; test the sign-inverting slope prospectively. An absent or wrong-signed inversion in well-measured cohorts is a falsifier for this operator-level prediction. Below are explicit, preregisterable predictions with corresponding measures, analysis plans, and falsifiers.

10.3.1 Core hypothesis (directional, not vague)

H1 — Valence inversion: Prior-day negative drift ($\Delta L_d < 0$) increases the probability that dream content and sleep architecture will show *counterweight* features that reduce load or increase repair probability:

Architecture: higher early-night SWA (load relief) and higher late-night REM density (repair rehearsal) than after neutral days.

Content: greater odds of reconciliation, mastery, safety, or sobriety themes in dreams, relative to neutral days.

H2 — Next-day carry-through: Dream *ReliefGain* and *RepairGain* features predict next-day reparative actions (calls, apologies, task closures, pain management, help-seeking), controlling for baseline mood, trait conscientiousness, sleep duration, and external stressors.

H3 — Horizon sensitivity: When perceived horizon H_t is short (e.g. a looming deadline or crisis), the inversion effects (in architecture and content) and the carry-through to behavior are *stronger*.

10.3.2 Quantities to measure (same-day → night → next-day)

Prior-day ledger drift (ΔL_d): Change in daily HCl from morning mean to pre-sleep mean. (Anchors for context include pain level, social exclusion events, feelings of failure or mastery, and effort expenditure that day – see Section 7.1.). In the full ledger, a day's contribution is the time-integral of HCl over the day; this morning-to-evening difference is a preregistered proxy for within-day drift.

Sleep architecture: *Sleep onset latency*; *SWA power* in the first cycle (z-scored within-person); *REM density* in the last two cycles; *REM latency* (time to first REM).

Dream content features: Blinded annotation scores for ReliefGain (e.g. presence of threat resolution, soothing attachment, regained competence in the dream) and RepairGain (e.g. apology rehearsal, plan formulation, confronting an avoided cue in the dream).

Next-day reparative actions: Time-stamped behaviors within 24 h indicating ledger correction (messages or calls sent, meetings scheduled, tasks completed, use of therapy or analgesia), coded from logs or EMA reports.

Perceived horizon: Evening and morning self-reports of time perspective (e.g. “How limited does your future time feel right now?”), plus binary markers for any *looming deadline* or *limited window* the person reports.

10.3.3 Primary statistical tests (preregister)

Architecture tilt: We fit mixed models to test whether prior-day drift predicts changes in sleep stages:

Model 1: $\text{SWA_cycle}_1 \sim \beta_0 + \beta_1 \Delta L_d + C + (1|\text{person})$.

Expectation: $\beta_1 < 0$ (more deep sleep after negative days; since $\Delta L_d < 0$ denotes a worse day).

Model 2: $\text{REM_density} \sim \gamma_0 + \gamma_1 \Delta L_d + \gamma_2 H^{-1} + \gamma_3 (\Delta L_d \times H^{-1}) + C + (1|\text{person})$.

Expectations: $\gamma_1 < 0$ (negative days \rightarrow more REM overall), and $\gamma_3 < 0$ (stronger coupling under short horizons; as H^{-1} increases, the REM increase per negative drift is larger).

Logistic mixed model:

$\text{Pr}(\text{RepairGain} > 0) = \text{logit}^{-1}(\alpha_0 + \alpha_1 \Delta L_d + \alpha_2 H^{-1} + \alpha_3 (\Delta L_d \times H^{-1}) + C + (1|\text{person}))$.

Expectations: $\alpha_1 < 0$ (more repair themes after more negative days), and $\alpha_3 < 0$ (the negative-day repair tilt is stronger when horizons are shorter).

Carry-through mediation: We examine the path $\Delta L_d \rightarrow \text{DreamFeatures} \rightarrow \text{Next-day Repairs}$. We use multilevel mediation (with bootstrap CIs) to test if dream Relief/Repair features carry part of the effect of a tough day on next-day reparative actions. The indirect effect should be positive. *SESOI*: for example, an odds ratio ≥ 1.25 per $+1\text{ SD}$ of RepairGain predicting a next-day repair action.

Specificity vs. simple mood repair: As a robustness check, we add next-morning mood as a covariate in these models. QS predicts that even after accounting for how you *feel* in the morning, the dream features retain unique predictive power for your reparative actions (i.e. it's not just "I felt better, so I did more").

10.3.4 Secondary predictions (fine-grained)

Valence inversion within-person: The same individual shows reliable inversion effects across multiple tough days (e.g. test-retest ICC ≥ 0.40 for that person's $\Delta L_d \rightarrow$ dream-feature slope).

Theme-selective inversion: Specific imbalances prompt specific counterweights: social rupture days preferentially yield reconciliation motifs; humiliation days yield mastery motifs; overindulgence days yield sobriety/exposure motifs in dreams.

Temporal placement: Relief-heavy content tends to occur in early-night (SWS-rich) episodes, while repair-heavy content clusters in late-night REM episodes (lining up with the physiological timing of relief vs. rehearsal).

Action friction: Next-day *decision latency* for reparative actions decreases after high-RepairGain dreams. (In cognitive terms, the starting-point bias in a drift-diffusion model shifts toward the repair option.)

10.3.5 Rival explanations and how we beat them

Hedonic adaptation: This classic theory predicts a general reversion to baseline mood, but not the *directionally appropriate dream content motifs* tied to the prior day's asymmetry, nor any horizon \times inversion interaction. (In other words, adaptation says "you bounce back eventually," but doesn't predict *how* dreams tilt or that time urgency matters.)

Generic consolidation: Standard memory consolidation theory predicts improved memory for experiences but not a policy tilt toward *compensatory* actions of equal immediate utility. If dreams were just reinforcing memories, we wouldn't necessarily see them biasing *which* actions people take.

Trait conscientiousness: Perhaps conscientious people both dream more about duties *and* tend to do more repairs. We will measure traits and control for them. QS's prediction is a *within-person*, state-dependent shift—if it were just a fixed trait effect, it wouldn't selectively amplify after tough days or with short horizons.

Sleep quality/duration: Better sleep generally leads to better mood and function. We will directly control for total sleep time and subjective sleep quality. QS predicts *specific*

architecture composition and content effects independent of total sleep. (It's not just "bad day = worse sleep"; it's how the sleep is structured and what is dreamed.)

10.3.6 Minimal viable protocol (community lab)

Participants: ~80 adults, ages 21–60, mix of chronotypes; each tracked for 28 days.

Tools: A validated wearable for sleep staging; a smartphone app for EMA prompts (4× per day); smart alarm capability for REM-targeted awakening ~2 nights per week; morning voice-journal for dream reports; automated logging of texts/calls/tasks for action tracking.

Blinding: Dream coders are blind to prior-day drift when scoring content; analysts are blind to condition order (days are coded); insert a few decoy "neutral" nights with dummy data to ensure coders aren't just guessing based on tone.

Power: Simulations should target detection of small effects ($\beta \approx 0.10\text{--}0.15$) with random slopes per subject. Preregister SESOIs and a stopping rule (e.g. stop only for poor data quality, not for peeking at results).

10.3.7 Field signatures for naturalistic datasets

Even without a lab study, we could look for these patterns in existing or opportunistic data:

Semester ends / product deadlines: In populations (students, teams) facing a fixed deadline, we'd expect to see *menu shrinkage* (people narrowing down to closure-focused actions) and a spike in reparative communications the day after REM-dense nights near the deadline.

Caregiver weeks: For individuals in caretaker roles, weeks with high evening stress should show more reconciliation-themed dream reports, and the following day an increased likelihood of help-seeking or respite-taking actions.

Public crises: At the city or region level, during public crises or disasters, we might see aggregate patterns – e.g. search query trends, hotline calls, or social media sentiment tilting in a compensatory direction – *lagging* one night after peak crisis days (consistent with a one-night "processing" delay via sleep).

10.3.8 Falsifiers (what would make us retract the claim)

No inversion: After controlling for confounds, ΔL_d fails to predict any compensatory tilt in architecture or content (dreams are affectively flat reflections of the day, not inverse counterweights).

No carry-through: Dream features do not predict next-day repair actions beyond what can be explained by mood or sleep duration alone.

No horizon interaction: The magnitude of inversion does not increase when horizons are shorter (no extra tilt under urgent conditions).

Rival sufficiency: A compact rival model (e.g. adaptation + overall sleep quality) fits the data as well or better than the QS-augmented model (e.g. $\Delta\text{AIC} = \text{AIC}_{\text{rival}} - \text{AIC}_{\text{QS}} \leq 2$, $\Delta\text{BIC} = \text{BIC}_{\text{rival}} - \text{BIC}_{\text{QS}} \leq 0$, Bayes factor ≤ 1 , no ΔWAIC improvement). In other words, adding Φ or inversion terms doesn't improve predictive accuracy.

Replicated evidence for any two of the above (under strict blinding) would force us to downgrade *valence inversion* from a QS signature to a non-specific sleep/mood effect.

10.3.9 Practical takeaway

On tough days, expect *cheap counterweights at night* and use them: capture your open loops before bed, protect your early-night deep sleep and late-night REM, and convert morning “repair-ready” feelings into one small, reversible action. For science, the point isn’t self-help; it’s that these predictable, directional patterns are there to be measured—or to fail—right now.

10.3.10 Where we go next:

Predictions demand careful sampling. Section 10.4 provides the practical playbook—REM timing, awakenings, coding schemes, and preregistered pipelines—so that dream data can be trusted.

10.4 Research Notes: REM Timing, Sampling, Coding

This section is a methods toolbox for capturing the “night workshop” effect without fooling ourselves. It covers when to sample REM, how to collect dream reports, how to code them into Relief/Repair features, and how to keep everything blind, reliable, and reproducible.

10.4.1 Timing: hitting the right REM

Why timing matters: Our theory predicts physiologic relief early (SWA-rich NREM) and repair rehearsal later (REM-dense bouts). To detect valence inversion, we must target *late-night REM* while also measuring early-night SWA.

Minimal timing plan (home or lab):

Baseline night: No experimental awakenings; establish each participant’s baseline SWA profile (cycle 1) and typical REM latency/density.

Experimental nights:

Early sample: Awaken once ~90–120 min after sleep onset *only if* deep NREM (stage N3) has occurred (to confirm a relief opportunity happened). Collect a brief state check (no full dream report unless the participant volunteers one).

Late REM sample: Awaken once in the last third of the sleep period during a sustained REM bout (see 10.4.2). Collect a full dream report upon awakening.

If lab-based: You can schedule the early vs. late awakenings on different nights (to reduce burden), e.g. one early-NREM awakening on night 1 and one REM awakening on night 3.

If home-based: Use wearables plus a phone alarm app guided by a REM-probability heuristic (see 10.4.2) to approximate the late REM wakeup.

Contingency: If a participant shows a clear REM rebound night (a spike in REM density) following a high-stress day, prioritize sampling that night’s late REM (since we expect the strongest signal then).

10.4.2 Staging and detection in practice

In-lab (gold standard): Use EEG/EOG/EMG to score 30-sec epochs by standard AASM rules.

REM detection: identify sustained low-amplitude mixed-frequency EEG with sawtooth waves (if present), plus phasic EOG bursts and muscle atonia on EMG.

Awakening cue: Use a gentle method (the person's name spoken softly or a 50–55 dB tone) aiming for a mid-REM awakening (avoid the very start of REM to reduce N1 contamination in the report).

At home (validated proxy):

Wearables: Use a device with accelerometer + PPG (heart rate) that provides estimated sleep stages as a guide.

Heuristic: Trigger the alarm when HRV dips and heart rate rises above the NREM baseline, movement is minimal, *and* the device has labeled the epoch as REM for ≥ 6 min continuously.

Fallback: If device staging is poor, use a fixed-time awakening in the last $\sim 20\%$ of the person's habitual sleep duration (e.g. if they usually sleep 8 h, wake them around the 6.5 h mark) as a simple proxy for late REM.

Quality checks: Log each night's sleep onset latency, REM latency, REM bout durations, and REM density (e.g. phasic eye movement counts per minute in lab; heart-rate surrogate at home). Exclude any awakenings that happened within ~ 2 min of a stage transition (to avoid confounding N1 or wake).

10.4.3 Eliciting reports without bias

The prompt (standardized):

"Please tell me everything that was going through your mind just before you woke up, in as much detail as you can remember, from beginning to end. If nothing comes, say 'no recall.' Do not interpret; just describe."

We give this exact prompt every time to avoid leading the witness.

Collection channels:

Lab: audio-record via a headset or bedside microphone, time-stamped to the awakening.

Home: a phone app with one-tap recording; encourage ~ 30 –90 seconds of narration (max 3 min) immediately upon waking.

Minimizing suggestion:

No mention of "repair," "forgiveness," or "mastery" (or any specific theory-related terms) until after data collection is complete. The participant should *not* know we're especially interested in those themes.

Do *not* ask “How did that dream make you feel?” on the first pass. First get the content; collect separate valence/arousal ratings afterward (simple sliders) so as not to bias the content report with emotional framing.

Recall control:

Log whether each awakening yields a recall or not, the word count of the report, and latency to begin speaking. These can be used as covariates later (some people or conditions might recall more, which needs to be accounted for).

Encourage an evening sleep diary (note if anything unusual or any substances, etc.) so that if recall is low, we can distinguish a true null from just “too sleepy to remember.” Also, writing things down at night can sometimes improve morning recall by reducing interference.

10.4.4 Coding: from stories to features

We convert messy dream narratives into numeric features that map to our ReliefGain and RepairGain constructs (plus a few auxiliary variables).

Feature dictionary (each scored 0 = absent, 1 = present in minor way, 2 = present and central):

Threat → Safety resolution (T→S): A threat or danger in the dream is explicitly resolved or escaped to safety.

Rupture → Reconciliation (R→R): An interpersonal conflict or separation is resolved—an apology given, forgiveness, positive reunion, or mutual understanding achieved.

Humiliation → Mastery (H→M): A scenario of failure or embarrassment flips to one of competence or success (e.g. you perform well or overcome the obstacle).

Indulgence → Sobriety/Exposure (I→S): A scenario of excess or avoidance turns into restraint or confrontation (e.g. refusing the substance, facing the feared cue).

Soothing attachment (SA): The dream includes warm, comforting contact with a trusted figure (non-sexual support, like a parent or close friend providing comfort).

Rehearsed closure (RC): The dreamer practices a real-life task or conversation (e.g. giving a speech, saying goodbye, confronting a boss) that could help achieve closure in waking life.

Reversibility cue (REV): The dream highlights options that keep future choices open (finding a safe exit, discovering a backup plan, “saving progress” in some way).

Intrusion down-weighting (ID): A previously intrusive or traumatic cue appears in the dream but is less potent or is normalized (e.g. the frightening figure from yesterday is now friendly or powerless).

Novel solution emergence (NS): The dream presents a new plan or solution not evident in the prior day's thinking (an insight or creative workaround appears in the dream narrative).

Mapping to ledger terms:

ReliefGain = $f(T \rightarrow S, SA, ID)$ – these features contribute to immediate load relief.

RepairGain = $f(R \rightarrow R, H \rightarrow M, I \rightarrow S, RC, NS, REV)$ – these features contribute to increased probability of waking repair actions.

(*Scoring*: Each feature is 0, 1, or 2 as defined. We sum the scores within each composite. We also predefine what we consider a meaningful change—e.g. +1 SD in these composite scores—as our SESOI for effects.)

10.4.5 Blinding and reliability

Blinding layers:

Coders of dream content are blind to the prior-day ledger scores, the horizon status, and the participant's identity. They only see anonymized dream texts.

Analysts are blind to condition order: they receive only feature tables and sleep metrics labeled with hashed IDs (so they can't tell if "Night 17" was after a bad day or a good day without the key).

Decoy nights: We insert some *neutral* dream reports (or even dummy text) randomly into the coding queue to monitor coder drift or hypothesis-guessing. Coders should not be able to tell real high-drift nights from decoys beyond chance.

Training and calibration:

Build a gold-standard library of ~50–60 example dream reports spanning clear cases of each feature. Train coders on these and calculate Fleiss' κ (agreement) and intraclass correlation (ICC) for the composite scores.

Aim for $\kappa \geq 0.70$ on each individual feature and $ICC \geq 0.75$ for the ReliefGain and RepairGain composite scores. If κ for a feature drops more than 0.10 on monthly re-checks, pause and recalibrate the coders on that feature.

Adjudication:

If two coders differ by a large amount (e.g. a score difference ≥ 2 on any key feature for the same report), it triggers a third coder to resolve. The final score is the *median* of the three. All such disagreements are logged for analysis (to see if certain features are inherently harder to judge, etc.).

10.4.6 Linking content to architecture

Tie each coded report to exact sleep metrics from the preceding bout and the night as a whole:

Bout-level metrics: time since sleep onset at awakening, sleep cycle number, REM density in that REM bout, any micro-arousals before the awakening.

Night-level metrics: SWA power in cycles 1–2, total REM percentage, REM latency, fragmentation index (number of awakenings or stage shifts).

Coupling tests: We hypothesize specific content-architecture links:

RC (rehearsed closure) and *R→R* (reconciliation) features should correlate with late-night REM density (within-person, across nights). In other words, when someone has especially dense REM, their dream is more likely to include closure or reconciliation elements.

SA (soothing attachment) and *T→S* (threat→safety) features should correlate with early-night SWA on nights following a negative day. High SWA (lots of deep sleep) earlier in the night might be associated with dreams that include comfort or threat resolution by morning.

Use mixed-effects models with random intercepts (and possibly slopes) per person to test these couplings (to account for individual baselines). These tests check if physiology and content are not just parallel effects but actually linked.

10.4.7 Minimal annotation software stack

Transcription: Use automated speech-to-text (ASR) to transcribe voice-recorded dream reports, with a human pass only for segments marked as unintelligible (e.g. tag “[inaudible]” where needed). This speeds up content availability.

Annotation interface: A web app or GUI that presents one dream report at a time with the locked rubric (the feature list with definitions and examples). Provide tooltips for each feature and one-click scoring (0/1/2) with an optional confidence slider (0–100%).

Versioning: Every time a coder makes or changes a score, create a new entry (with timestamp, coder ID, changes). This ensures a complete audit trail of how scores evolved (and discourages post-hoc “fitting”).

Exports: Data should export to JSON and CSV with a clear schema: each row contains a unique report ID (or hash), feature scores, any confidence ratings, and optional free-text notes or comments.

10.4.8 Quality controls and exclusion rules

Exclude dream reports that have fewer than ~10 words *unless* the coder explicitly marks it as “vivid but concise.” (Reason: extremely short reports usually indicate poor recall and aren’t reliable.)

Flag awakenings that show signs of stage-transition contamination (e.g. EEG indicates the person was actually drifting out of REM into N1 at awakening). We might exclude or analyze separately, since those reports could be muddled.

Control for major confounds via data flags: e.g. nights with sleep medications, alcohol use after 2 pm, or acute illness. These are recorded and later included as covariates or used to exclude certain nights if necessary. (We preregister how we’ll handle these.)

Monitor device data quality: require that $\geq 70\%$ of nights have valid sleep staging data. If a participant’s wearable fails often (below this threshold), we either use only their self-reports or drop them, depending on plan.

10.4.9 Power and sampling notes

Expect small effects per night (e.g. an odds ratio in the ballpark of 1.2–1.4 for a dream feature predicting an action). We rely on within-person designs to boost power by controlling individual differences.

Rule of thumb: Aim for at least ~60 participants \times 14 nights each \times ~6 usable REM-dream reports each. This yields a robust dataset for estimating person-specific slopes ($\Delta L \rightarrow RepairGain$ and $\Delta L \rightarrow ReliefGain$) with enough precision. (In simulation, this gave stable random-slope estimates.)

We use simulation-based power analysis with the planned analysis pipeline (state-space model for HCI, mixed models for dream effects, etc.) to confirm required N, explicitly incorporating uncertainty propagation from $\hat{L}(t)$ estimates and within-person variance. Adjust sample size or duration accordingly.

10.4.10 Pre-registration checklist (what to lock before data)

To avoid *p-hacking* or hindsight bias, we lock in key analysis decisions:

Primary outcomes: Dream *RepairGain* and *ReliefGain* composite scores; REM density; SWA power in cycle 1.

Predictors: Prior-day ledger drift (ΔL); perceived horizon (and its operationalization); the $\Delta L \times$ horizon interaction term.

Covariates: Age, sex, chronotype, total sleep time, flags for caffeine/alcohol after cutoff, relevant medications. (We decide these a priori.)

SESOI: e.g. $\beta \geq 0.10$ for the effect of ΔL on REM density; OR ≥ 1.25 per +1 SD *RepairGain* for predicting a next-day repair action. These thresholds define what we consider meaningful.

Missing data plan: Use multiple imputation for intermittent missing EMA entries; do sensitivity analysis to see if results change under worst-case assumptions (MNAR).

Stopping rules: No peeking at outcomes. Interim looks only check data quality (are people complying? is the device working?), not effects. We commit to a fixed sample or a predetermined sequential analysis plan.

10.4.11 Ethics notes

Limit dream-sampling awakenings to ≤ 2 per week and ≤ 6 total per participant across the study. We don't want to ruin anyone's sleep.

Allow *skip nights* or rescheduling if participants have critical obligations (exam next day, sick child, etc.) or are ill/exhausted. Participant well-being > data.

Provide every participant with personalized sleep feedback and (if they want) a summary of their own dream patterns after the study. Regardless of results, they should gain something positive (better sleep habits, insight, etc.) from participating.

Relief is a systems variable, but a person's comfort and dignity override data collection. We only study what we can do with minimal intrusion and maximum compassion.

10.4.12 What would invalidate your pipeline

It's important to define not just what would invalidate the *theory*, but what would indicate our *methods* are failing:

Stage mislabeling: If our home REM detection (wearable + heuristic) is wrong > 15% of the time compared to lab PSG, we have to either repeat critical parts in the lab or improve the algorithm. Otherwise, we might be sampling the wrong thing entirely.

Coder bias: If coders, despite blinding, are *somewhat* able to guess prior-day drift from the dream reports at above-chance rates (meaning our content coding isn't truly blind), we need to revamp the prompts and retrain coders. We might add more decoys or mask emotional tone in transcripts.

Null mediation under power: If with a decent N and proper controls we find that Relief/Repair features do *not* mediate the link between tough days and next-day behavior (the indirect path is effectively zero), then dream features might be epiphenomenal. We would have to treat the whole dream-content angle as unsupported and downgrade those claims (see 10.6).

With good REM timing, clean prompts, blinded feature coding, and preregistered models, we can turn dreamy anecdotes into auditable data. If QS is truly using the night to rebalance the day, these methods are sharp enough to catch it—or to show that the pattern just isn't there.

10.4.13 Where we go next:

Methods in hand, Section 10.5 gets personal: a one-week dream ledger exercise that classrooms and clinics can try, carrying uncertainty and respecting privacy at every step.

10.5 A One-Week Dream Ledger Exercise

This is a self-run, low-burden protocol you can try at home to glimpse how sleep may rebalance your day. It won't prove LoF on its own, but it will show you how to track a ledger, sample your dreams, and convert night signals into small morning repairs. In just seven days, you can collect enough data to see whether a valence-inversion pattern appears in your own life.

10.5.1 What you need

A phone (set to airplane mode at night) or a small notebook by the bed.

A timer or alarm app. (If you have a validated wearable that tracks sleep stages, even better—but not required.)

A simple daily form (you can draw a table or use a notebook page) to record:

Morning HCI (0–10): How you feel overall upon waking (happiness/comfort composite).

Evening HCI (0–10): How you felt overall in the afternoon/evening.

Anchors (0–10 each): Quick ratings of key contributors: Pain, Social Friction, Mastery/Competence, Fatigue/Effort (each 0 = none, 10 = extreme).

Open loops: 1–3 unresolved items that matter for tomorrow (e.g. “need to call X,” “worried about Y”).

Morning actions: Did you do one small “repair” action this morning? (Yes/No, and what was it?)

10.5.2 Daily schedule (7 days)

Evening (≈5 minutes): Rate your Evening HCI and the four anchor items. Write down 1–3 *open loops* you'd like to close soon (a call to make, an apology, a form, a chore—anything weighing on you). *Optional:* Place a neutral object (a certain scent or a soft sound) by your bedside as a cue; you'll use it upon waking.

Night: Go to sleep as usual. If you're using a wearable, no special action is needed. If not, just sleep — we'll rely on your natural waking. (Set an alarm for your usual wake time if you must wake up by a certain time, otherwise let yourself wake naturally once.)

Upon waking (within 3 minutes): Before moving much or starting your day, capture what was in your mind *just before* you woke. Speak into your phone's recorder or jot down 5–10 sentences in your notebook. Don't interpret the dream or story—just document it. On a new line, rate the Dream Affect from –4 (very unpleasant) to +4 (very pleasant) and

Dream Vividness from 0 (not at all) to 4 (extremely vivid). Now scan your dream story and tick any of the following features if present: Threat→Safety, Rupture→Reconciliation, Humiliation→Mastery, Indulgence→Sobriety/Exposure, Soothing Attachment, Rehearsed Closure, Reversibility Cue, Intrusion Down-weighted, Novel Solution. Finally, choose one tiny, reversible morning action that moves one of your open loops forward by one step (send a text, schedule a call, tidy for 5 minutes, drink a glass of water and set up an appointment, etc.).

Midday (≈ 1 minute): Mark whether you completed that one small action by noon. If not, jot a word or two about why not (no judgment—just data).

Night (≈ 2 minutes, end of day): Use your Morning HCl rating from waking and your Evening HCl rating. This now lets you compute a simple daily drift: $\Delta L_d \approx$ Evening HCl – Morning HCl. (Negative means the day was tougher than the morning; positive means the day improved.)

10.5.3 A pocket scoring rubric (3 minutes each day)

To quantify your dream quickly:

Go through your dream report and give a 0 or 1 for each feature on the list if it was present (1) or not (0). Specifically:

Relief features: Threat→Safety, Soothing Attachment, Intrusion Down-weighted.

Repair features: Rupture→Reconciliation, Humiliation→Mastery, Indulgence→Sobriety/Exposure, Rehearsed Closure, Novel Solution, Reversibility Cue.

Compute two quick composite scores:

ReliefGain = sum of relief feature scores (range 0–3).

RepairGain = sum of repair feature scores (range 0–6).

(*If a feature was strongly present, you can count it as 1; if not at all, 0. Since this is informal, we won't worry about “2 = central” for now.*)

10.5.4 What to look for in your own data

Now you have a miniature dataset. Here's what to check after 7 days:

Valence inversion: On *tough* days (when ΔL_d is negative, meaning evening mood < morning mood), do you see higher ReliefGain or RepairGain in the *next* morning's dream compared to easier days? In other words, did bad days tend to be followed by more “positive” or compensatory dream content?

Morning carry-through: Do higher RepairGain scores go along with a higher chance that you *actually did* the one small repair action that morning? (For example, on mornings after dreams full of reconciliation or mastery themes, did you more often complete your planned action?)

Horizon effect: If one of your open loops had a real deadline or limited window (say something *had* to be done the next day), check the prior night's dream: did it show more targeted repair themes? And did that morning feature more decisive steps? In contrast, when nothing urgent was looming, dreams might not tilt as much.

If you happen to use a wearable and notice a “REM-rich” night after a particularly hard day, check whether that night’s dream had strong repair motifs *and* whether you got your one-step repair done faster or more readily than usual. This is anecdotal, but it’s exactly the pattern we’d predict.

10.5.5 A seven-day worksheet (template)

You can organize your data as a one-line-per-day log: Day; ΔL_d (Evening–Morning); ReliefGain (0–3); RepairGain (0–6); Repair before noon? (Y/N); and a brief note on the action.

Day	ΔL_d (Evening–Morning)	Relief-Gain (0–3)	Repair-Gain (0–6)	Repair before noon? (Y/N)	Note the action
1					
2					
3					
4					
5					
6					
7					

Mini-insight rule: If ≥ 4 of your days with negative drift are *followed* by higher Relief/Repair scores in dreams *and* at least 3 of those mornings include a completed action, you’ve observed a personal valence-inversion signature. (If not, that’s useful data too!)

10.5.6 Troubleshooting and tips

No recall? If you don't remember a dream on a given morning, just write "no recall" and move on. (Dream recall often improves by day 3 as you get used to this.) You can still record how you *felt* upon waking and do the one-step action.

Time-poor mornings? Do an ultra-micro repair. For instance, if you planned to write an email but have no time, send a two-sentence text or even fill out just one line of a form. The key is to do *something* reversible that addresses an open loop.

Anxious or upsetting dream content? Remember, you're not being asked to interpret or "fix" your dreams. Just notice the features and see if there was any silver lining (a twist toward safety or a solution). If a dream was disturbing, you can still identify if, say, it ended a bit better than it began. And regardless, you still take a small positive step in the morning.

Missed a day? Skip it and continue. Don't try to backfill or catch up by doing extra — consistency is more important than perfection.

10.5.7 Privacy, ethics, and boundaries

Protect identities: Don't record real names or identifying details in your dream notes. Use initials or roles ("my sister," "my boss").

Emotional safety: If a dream's content is disturbing or traumatic, it's okay to note "disturbing content present" without details. And if it brought up serious distress, consider talking to someone (friend or professional). This exercise is not meant to push into trauma processing.

This is not therapy: Treat this as a self-observation exercise, not a clinical intervention or a way to *force* positive outcomes. Its purpose is to illustrate a concept and gather data, not to solve your life in a week. Be kind to yourself no matter what the data say.

10.5.8 Optional enhancements (for the curious)

Evening cueing: As mentioned, you can pair a neutral scent or soft sound with reviewing your open loops at bedtime, then present the same cue when you wake up. The idea is to *subtly bias* recall and perhaps the dream content (e.g. a rosemary scent or specific tone might weave into your dream). This is a gentle form of TMR and can sometimes increase the continuity between your intention and dream.

Accountability buddy: If you want extra motivation, share *only* your one-step morning action (not your whole dream unless you want to) with a friend or partner each morning,

and have them share theirs. Knowing someone expects a text (“What was your small action today?”) can boost completion rates.

Week 2 swap: If you do a second week, try a *planned intervention*: e.g. schedule a low-stakes reconciliation midweek (call someone or resolve a minor conflict on purpose) and see whether the previous night’s dream carried any practice motifs (perhaps your brain anticipates it). This is just for exploration.

10.5.9 What would count as not seeing it

Flat line: Across your seven days, tough days show no increase at all in Relief/Repair features and no rise in morning repairs compared to easier days. (Dreams just mirror your days, or vary randomly.)

Opposite pattern: Tough days are *followed* by dreams that are even more negative (indulgence or avoidance motifs) and you feel even less inclined to repair in the morning.

Randomness: The features and actions vary without any relation to your ΔL_d . No pattern — good days sometimes have compensatory dreams, sometimes not; same for bad days.

If you observe these “null” or contrary patterns over multiple weeks, your personal data do not support the “night workshop” claim. And that’s genuinely useful to know! It’s exactly the kind of negative evidence science needs to refine theories. It might mean any number of things — e.g. maybe the effect is real but only shows in certain people or over longer periods, or maybe dreams aren’t doing what we think. Either way, you’ve contributed by testing it.

10.5.10 Why bother?

Because LoF makes a sharp prediction: when the day tilts, the night leans back. This one-week exercise lets you see whether that lean shows up in your own ledger and dreams—no lab, no jargon, just a ledger, a pen, and seven days of honest noticing. Either you’ll spot the pattern, or you’ll gather evidence that *challenges* it. Both outcomes push our understanding forward.

10.5.11 Where we go next:

Exercises must face failure modes. Section 10.6 spells out what patterns would count against the dream-counterweight idea and how to report nulls without spin.

10.6 Fail Patterns in Dream Data

This chapter's claims live or die on directionality, carry-through, and horizon sensitivity. Below is a concrete failure taxonomy—patterns that, if replicated under blinding and preregistration, should make us downgrade or abandon the “night workshop” account of dreams.

10.6.1 Directionality failures

Flat directionality: Prior-day negative drift does *not* predict *any* increase in Relief/Repair dream features; coefficients hover near 0 with tight confidence intervals. (Dream content is unrelated to yesterday’s imbalance.)

Wrong-way inversion: Tough days are actually followed by *more* indulgence or avoidance motifs (instead of reconciling/mastery) in dreams — a stable opposite trend across samples.

Theme non-specificity: Specific ledger imbalances don’t yield matching counterweights. Social-rupture days do *not* preferentially lead to reconciliation themes; humiliation days do not yield mastery dreams, etc.

Consequence: We would downgrade QS from a targeted counterweight mechanism to just generic consolidation. In practice, we’d stop claiming “valence inversion” and acknowledge that dreams aren’t reliably compensatory.

10.6.2 Carry-through failures

No behavioral mediation: Dream features fail to predict next-day repairs after controlling for morning mood, sleep duration, and personality. (Dreams had no unique influence on behavior.)

Latency unaffected: The decision latency for reparative actions (how quickly you act) is no shorter after high-RepairGain dreams. Dreams don’t reduce any friction to acting.

Substitution effect: Dream features correlate with self-reported *insight* or catharsis, but not with any objective actions (e.g. someone feels like the dream “taught a lesson” but doesn’t actually do anything differently).

Consequence: We would have to treat dream content as *epiphenomenal* — nice stories, maybe therapeutic in feeling, but not driving action. The claims would be restricted to architecture-only recovery (e.g. “deep sleep helps recovery” without any policy/choice tilt).

10.6.3 Horizon insensitivity

No H^{-1} interaction: The compensatory effects do not get stronger when horizons shrink. Short-term crises or end-of-life scenarios show no extra dream inversion compared to normal times.

Uniform menus: Proxies for the admissible set (like morning breadth of options or “menu width”) do not narrow even when time is obviously short.

Consequence: QS loses its signature of *tightening constraints under shrinking horizons*. We’d have to say dreams (and by extension LoF) behave like adaptation or homeostasis, unaffected by how much time is left. The whole “urgency” aspect would be dropped.

10.6.4 Architecture–content decoupling

REM density unlinked to RepairGain: Within the same person, nights with peak REM density don’t carry more repair motifs in dreams. The physiological “extra REM” might happen with no corresponding narrative work.

SWA unlinked to ReliefGain: High early-night SWA fails to predict any next-morning load reduction or soothing content. The deep sleep doesn’t translate to comforting dream elements or calmer wake-ups.

Stage misattribution: The supposed effects appear even when dreams are actually N1/N2 or wake imagery. For instance, if we accidentally count light sleep reports as REM and still see “effects,” it means the REM-specific claim is on shaky ground.

Consequence: We would consider our architecture claims likely artifacts of devices or chance. We’d conclude maybe dream *content* has some psychology but the physiological angle (SWA/REM as LoF tools) isn’t valid. So we’d restrict conclusions to “what people dreamed” and drop “because REM did X.”

10.6.5 Reliability and blinding breakdowns

Coder drift: Our feature scoring reliability falls apart over time (e.g. $\kappa < 0.60$ after a while) or coders start unconsciously inferring the person’s day from the dream (blinding fails).

Analysis leakage: Analysts (or algorithms) end up unblinded to which days were “bad” vs. “good,” and once you enforce true blinding, the effects shrink to null. (This would mean some inadvertent peeking or p-hacking occurred.)

Recall confounds: Simple metrics like word count, recall success, or dream vividness fully explain what we thought were dream-feature effects. For example, perhaps on bad

days people just *remember more dreams* (due to lighter sleep), which gives more “content” to find patterns in.

Consequence: Invalidate the dataset. We’d have to rerun with stricter blinding, tighter prompts (to avoid leading), and include recall/vividness as covariates. Basically, we’d say “our initial data were too messy to trust.”

10.6.6 Rival sufficiency

Adaptation + sleep quality wins: A lean model that includes just baseline mood regression plus total sleep time and a stress index matches or beats the full QS model on all metrics (e.g. $\Delta\text{AIC} = \text{AIC}_{\text{rival}} - \text{AIC}_{\text{QS}} \leq 2$, $\Delta\text{BIC} = \text{BIC}_{\text{rival}} - \text{BIC}_{\text{QS}} \leq 0$, Bayes factor ≤ 1 , and no better predictive log-loss for the QS model). In plain language, adding all our additional LoF terms doesn’t improve predictions.

Trait model dominance: Individual differences (like trait conscientiousness or agreeableness) explain who does morning repairs, and adding dream features changes R^2 by <0.01. Essentially, it was just personality all along.

Consequence: Prefer the rival explanation. We’d conclude the LoF/QS additions are unnecessary complexity. The Law-of-Fairness dream hypothesis would lose out to the simpler theory that people just revert to baseline or act based on stable traits.

10.6.7 Negative controls that flip or go null

Decoy nights “predict” repairs: If we insert completely neutral or random dream reports as decoys, and those *also* predict next-day actions as well as the real REM reports do, then our supposed effects are not real. (Maybe our coders or analysts are fooling themselves.)

Time-shift test fails: If prior-day drift correlates just as strongly with *two days later* behavior as it does with next-day, it implies any link we saw might be an artifact (like weekly cycles or people having patterns every other day) rather than a causal nightly reset.

Benign days, “repair” dreams: If days scored as positive (no issues) are followed by *equally high* RepairGain dreams *without* any corresponding need or action, then our dream coding might be picking up something generic (like imaginative people always dream with lots of content) rather than true counterbalancing.

Consequence: We’d suspect expectancy effects or some subtle weekly structure rather than our hypothesized mechanism. We’d have to remove any claim of *directional* compensation via dreams.

10.6.8 Dose-response and specificity failures

No dose-response: Larger negative drifts (really bad days) do *not* produce stronger inversion effects than mild negative drifts. The slope of “worse day → more dream compensation” is flat.

No theme specificity: The content of the day’s pain doesn’t match the content of the dream’s remedy. For instance, days heavy in physical pain don’t preferentially yield soothing/relief content; days heavy in social friction don’t preferentially yield reconciliation dreams.

Physiology non-specific: We see REM rebound nights more after random things like late caffeine or alcohol than after high-stress days. (So maybe REM rebound is just homeostatic, not tied to affect at all.)

Consequence: Abandon the content-specific predictions. We might still say “sleep helps regulate mood” but not in a targeted LoF way. Physiologically, we’d interpret extra REM or SWA as general recovery processes, not guided by a fairness ledger.

10.6.9 Cross-modality discordance

EMA vs. sleep split: Suppose our EMA (experience sampling of mood) shows that people do more next-day repairs after tough days, but our sleep metrics show no coupling to those repairs. That means people might be compensating *behaviorally*, but it has nothing to do with dream content or specific sleep stages.

Neural nulls: In the lab, if we measure QS-related brain signals (e.g. a proposed “QS-residual” in vmPFC or ACC that should indicate compensation need) and those vanish once we control for basic factors (utility, conflict, arousal, recall), then the neural signature of QS isn’t there. In other words, the brain data don’t back up an active counterweight mechanism beyond known processes.

Consequence: Drop the whole QS-residual idea (Φ in the brain) for dreams. It would suggest any fairness-like effect is happening in daytime behavior or subjective feeling, but not showing up in sleep physiology or neural measures. We’d confine LoF testing to daytime outcomes and end-of-life studies, where we might have better signals.

10.6.10 Cultural and symbol drift traps

Feature instability across cultures: Our Relief/Repair coding scheme might fail scalar invariance across cultures. Maybe coders in another culture can’t agree on what counts as “reconciliation” motif, or the frequency of features differs in ways that don’t map to

actual compensation differences. If the κ for features collapses outside the culture we designed it in, that's a problem.

Age-stratified reversals: Perhaps adolescents show the *opposite* patterns of older adults even with similar data quality, once you control for sleep need. Maybe teen dreams function differently (or their baselines differ so much that our scoring doesn't translate).

Consequence: We must restrict generalization. It could be that the "night workshop" is a culturally specific phenomenon or requires culture-specific rubrics. We'd require configural→metric invariance checks whenever comparing groups. At minimum, we'd say we need different coding schemas or extra calibration for different ages/cultures, and we'd be cautious in claiming universality until that's resolved.

10.6.11 Device and staging artifacts

Wearable mis-staging: If our home wearables disagree with gold-standard PSG on > 15% of REM bouts, and crucial effects *vanish* when using lab-verified data, then any findings might have been artifacts of bad data. (E.g. maybe the wearable mislabeled quiet wake as REM and we "found" something that isn't real.)

Transition contamination: If the effects we see (like dream content correlating with something) only appear when awakenings happen near stage transitions (i.e. when dreams were likely mixed with waking or N1), and disappear when awakenings are cleanly in middle-of-REM, then our data might be confounded by partial awakenings or lighter sleep moments.

Consequence: Treat the at-home staging as insufficient. To salvage the hypothesis, we'd have to rerun key tests with full PSG, or at least improve the algorithm significantly. Basically, we'd blame the gear and not make any strong claims until we had cleaner sleep staging.

10.6.12 Statistical red flags

Researcher degrees of freedom: If significant effects only appear after *post hoc* choices—like if someone bins the data just so, or drops one outlier, or includes an extra covariate—and otherwise nothing, that's a huge warning. (E.g. if we had to cherry-pick how to measure "tough day" to get any result.)

Winner's curse: Initial studies show large effects, but when we try to replicate with proper preregistration and larger samples, the effects shrink to nearly zero. This suggests the first finding was an overestimate (common in science).

Non-replication across labs: If one lab finds something, but two others do the same protocol and find nulls, and we can't identify any meaningful difference between labs, then it might be that the original was a fluke or results aren't robust.

Consequence: We put the brakes on theoretical claims. No more sweeping statements until multi-lab collaborations are done. We might design an adversarial collaboration where multiple teams test the hypothesis together. Essentially, we'd say "Pause on this theory until we figure out why these inconsistencies happened."

10.6.13 Stop-rules for the thesis (dream chapter)

We agree to *withdraw* the "night workshop" mechanism as a key LoF claim if, after two independent preregistered multi-lab studies:

Directionality and carry-through are both null (or worse, in the wrong direction) under rigorous blinding, and

A compact rival model explains the data as well as our model (equal predictive power with fewer parameters, meeting the criteria above for ΔAIC , Bayes factor, etc.).

At that point, the LoF research program would shift emphasis to other signatures (like end-of-life neutrality, horizon manipulation in waking behavior, etc.), treating dreams as *non-diagnostic* for QS. In short, if dreams don't deliver the evidence, we won't keep chasing them beyond that point.

10.6.14 What to report when it fails

Publish the null results or counter-results openly, with full code and the preregistered SESOI thresholds, and even the de-identified dream feature libraries if possible. Hiding failed studies only hurts science.

Provide an analysis of *why* it failed: for example, did recall problems, staging errors, or coder drift contribute? Include results of negative controls (they help show if something systematic went wrong or not).

State explicitly which LoF claims are retained, which are weakened, and which are relinquished in light of the results. Maybe the overall LoF is still alive but the dream part isn't; we need to say that clearly.

10.6.15 Where we go next:

With nights addressed, we move to life's closing days. Chapter 11 asks whether, under humane observation, we can see the fairness signatures that should be strongest near the end—always with comfort and dignity first.

Chapter 11 — End-of-Life: Where the Law Shows Its Hand

Some phases of life make our measurements murky; others bring them into sharp focus. The end of life is the sharpest. As a person’s horizon contracts, options narrow and trade-offs become stark. If the Law of Fairness is more than just a hopeful tendency—if it is a true constraint—its signature should intensify as death approaches. In plain language, the claim that “every conscious life’s ledger balances by the end” would have to really show itself when a life is at its end. This chapter lays out the straightforward logic, the humane study designs, and the specific observations that would signal the law “showing its hand” (or failing to) when conscious life draws to a close.

To appreciate why the end-of-life stage is so critical, imagine two very different final chapters. In one scenario, a dying person spends their last days in turmoil—unmanaged pain, unresolved conflicts, despair—leaving life with a sense of profound imbalance. In another scenario, a dying person, even after much suffering, experiences a late peace—pain is eased, goodbyes are said, old wrongs are forgiven—and they pass with an uncanny tranquility. Where is the fairness? If LoF holds, we expect the second scenario to be more than wishful thinking or good hospice care; it should be *predictable*, a natural consequence of the system pushing for closure. The end-of-life is where *either* the fairness law steps up (guiding the system toward neutral closure) *or* it doesn’t (and some lives truly end unfairly). There is little room for ambiguity here, which makes it a stringent test.

From the core LoF claim (every unified stream ends neutral at the “death of mind”), we derive three observable predictions as the end draws near:

1. Horizon scaling: As the expected remaining time H_t shrinks, an individual’s set of admissible actions should *tighten and tilt toward closure*. The person’s choices increasingly focus on things like saying goodbye, setting affairs in order, seeking meaning or reconciliation, and maximizing comfort. They will be less interested in starting new ventures or feuds. In short, when time is short, closing acts (resolving loose ends) crowd out open-ended ones.
2. Compensatory intensification: Counterweights that were diffuse earlier in life become more concentrated near the end. LoF predicts bursts of relief or clarity—for example, sudden rallies of lucidity, final moments of connection or forgiveness, or waves of calm—*provided the channels for them are open*. These aren’t miracles or supernatural events, but the natural intensification of any remaining balancing moves. What might have been subtle balancing efforts

earlier (a slight mood lift after a hard month) could manifest as a dramatic moment of peace or resolve when one's days are truly numbered.

3. Neutral closure: The net felt balance over the final stretch of life (say, the last few days or week of consciousness) should drift toward an emotionally neutral state. That doesn't mean euphoria or a sugar-coated ending. It means that, on comprehensive measures of affect (like our HCl), the person's mood in their final conscious window would be closer to the center (neither extreme high nor extreme low) than one might predict from their prior suffering alone. In other words, even someone who suffered greatly should, if given proper care, exhibit a leveling off – not a continued plummet – in those final days.

It's vital to clarify what "neutral" does and doesn't mean here. Neutrality at life's end does *not* imply that people die happy, or in denial, or that a tragic situation is suddenly "okay." It simply means that on a cumulative basis, the person's felt experience in the final conscious period isn't overwhelmingly positive or negative, but somewhere near the middle. For example, someone might not be joyful, but they might be surprisingly calm or emotionally even, given the circumstances. LoF's claim isn't that everyone dies with a smile; it's that no one dies with an unsettled debt of suffering or pleasure on their ledger. If a life has been pain-heavy, end-of-life should bring proportionate relief or uplift. If a life has been unusually easy, end-of-life might bring challenges or downs that temper that surplus. And if relief or challenges are *prevented* (say a person is in unmanageable pain or isolation), then the neutrality might not be reached—not because LoF is false, but because the mechanism was blocked. This last point shows LoF is a *constraint*, not magic: it can operate only when the conditions allow (we'll revisit this as a boundary condition).

Before discussing evidence, ethics come first. All end-of-life research must rest on unyielding ethical principles. We will not—and need not—do anything to *experiment* on dying individuals beyond compassionately observing what good care already does. Here are the non-negotiables we adhere to:

- Consent and assent: We obtain the patient's informed consent whenever possible. If a patient is unable to give full informed consent (due to cognitive impairment, etc.), we require surrogate consent from a legal guardian *and* the patient's assent (which could simply be a calm nod or any affirmative signal). The patient (or surrogate) can withdraw from the study at any time, no questions asked.

- Care primacy: Comfort, symptom control, and privacy always outrank data. We will never ask for a measurement or questionnaire if it in any way interferes with the patient's comfort or chosen focus. Research staff are essentially flies on the wall; the hospice or palliative care team's routine remains paramount. Think of our study as something layered gently atop standard care, never redirecting it.
- Minimal intrusion: All measures are designed to be as brief and unobtrusive as possible. For example, if we use any device, it might be a soft wristband that measures heart rate – nothing that beeps or requires blood draws or interrupts a patient's sleep. If we ask questions, it might be via a nurse who's already checking in, and it will be phrased in a yes/no or 0–10 format that takes seconds. Importantly, if a patient is resting or doesn't wish to respond, we skip it. Dying is not the time to fill out surveys unless the patient *wants* to share.
- Return of value: Whenever feasible, we give something back to the patient or family. This could be as simple as a summary sheet of the patient's comfort levels and alertness over the past week (which might help in care decisions or for family understanding). Or perhaps a recorded "legacy message" for loved ones, if the patient wants, facilitated by the study. We ensure that participation, however light-touch, yields some benefit or at least a kind of companionship, rather than feeling like an extraction of data.

In short, we only study what *good hospice care would already be attending to* – we simply measure it carefully and respectfully. If LoF has any reality, it should reveal itself through the very same channels that hospice and palliative practices aim to open: pain relief, social connection, emotional expression. We won't be creating any new interventions, just observing and analyzing what unfolds when those best practices are in place.

So, what can we actually measure, humanely, at end-of-life? We focus on low-burden, opt-in observations that integrate seamlessly with care:

- We might use a short-form Hedonic Composite Index (HCI) – say a 3- or 4-item mood and comfort survey (e.g. rating pain, ease of breathing, peace of mind, and connectedness on simple 0–10 scales) given three times a day. These item ratings are coded so that higher values always mean better experience and then centered (e.g., by subtracting the midpoint) so that 0 denotes neutral before forming the composite. Patients could self-report if able; otherwise, a caregiver or nurse could give a best-approximation. It's quick (under a minute) and uses simple language or even emoticon faces.

- A Relational Contact Log would note each day who visited or called, and whether any significant reconciliations or heartfelt conversations occurred. This is often documented informally in hospice; we'd formalize it. It tells us about social closure activities.
- Key Symptom and comfort data come for free via medical records: we'll track medication doses (pain meds, anxiety relief, etc.), vital signs related to comfort (like oxygen requirement or restlessness), and any notable events (e.g. a sudden agitation episode or a calm, lucid interval). This helps us quantify how well symptoms are controlled.
- Narrative markers: With permission, we might have a clinician or family member jot down brief notes when the patient says or does something that clearly relates to closure or acceptance (for example, expressing forgiveness, gratitude, or a life regret, or rallying to give advice to family). These qualitative notes, coded carefully, can indicate psychological and emotional resolution processes.
- Dreams or visions: Only if the patient volunteers and is comfortable, we would invite sharing of any vivid dreams or end-of-life visions. Some hospice patients report meaningful dreams or visions (deceased relatives, religious imagery, etc.). We would approach this very gently – perhaps the clinician simply notes if the patient mentions a dream or vision, and whether it seemed comforting or distressing. (This ties back to Chapter 10: even at end-of-life, LoF predicts some dreams might serve as powerful counterweights or preparation for closure.)
- Passive comfort metrics: If a patient is open to it, we could use a nonintrusive wearable (like a fitness tracker) to monitor things like heart rate variability (HRV) as a proxy for stress, or sleep patterns at night. But this is entirely optional and guided by the patient's tolerance. Some patients might enjoy knowing their sleep quality, for instance; others won't, and we won't push it.

Every measurement is opt-in and *customizable*. Patients can choose which, if any, devices or questions they're comfortable with. The guiding philosophy is: measure only what a person would not mind (or might even appreciate) being measured, and nothing more.

Now, what does LoF specifically predict in this context of end-of-life, and how will we recognize it? Let's crystallize the predictions into a set of testable statements:

- P1: "Menu" tilts toward closure. As H_t (perceived remaining time) becomes very small, we expect to see patients initiating more closure-oriented acts. This means, for example, a rise in things like making amends, saying goodbye, giving

final instructions or gifts, or other “last things” – and a corresponding drop in starting anything new or frivolous. We’d measure this via the contact log and narrative notes (how many closure acts per day? how many new disputes or new projects?).

- P2: Relief becomes a priority (when possible). Provided pain and symptoms are reasonably managed (because uncontrolled pain can mask everything), patients should increasingly report moments of calm, contentment, or even gratitude amid the decline. LoF suggests the system biases toward relief when it can. So, we might see HCl mood ratings that, despite physical decline, hover at a mild, neutral-positive level rather than tanking. We might also see patients taking more pleasure in small things (sunlight, a favorite song) – a sign that the system is squeezing whatever relief is available from remaining channels.
- P3: Last-window centering of mood. Quantitatively, the average HCl score over, say, the final 72 hours of consciousness should be closer to zero (neutral) than one would expect by extrapolating the person’s prior weeks or months. For example, if someone had been rating their days as 2/10 (very bad) consistently due to illness, a naive extrapolation is they’d die at 2/10. Neutral is around 5/10 on that kind of raw scale (which corresponds to ~0 on our centered HCl). We predict that with good care, that person might actually average, say, 4–6/10 in the final days – an upward drift toward the middle. It’s not *happy*, but it’s notably higher than expected. This is something we can test with statistical models (like checking if there’s a significant upward drift after accounting for symptom control).
- P4: Counterweight bursts (when channels open). We anticipate brief episodes that pack a lot of positive or meaningful experience into a short time. These could be moments of lucidity (a sudden clear, alert period where the patient interacts vibrantly despite prior confusion), or emotional breakthroughs (like a short but profound conversation or expression of love/forgiveness), or even a painless window where the person is unusually comfortable and talkative. LoF frames these as the system delivering “compensatory payouts” when conditions allow (pain managed, loved ones present). We’d detect these in our data as spikes or outliers of high HCl or high interaction/closure activity, correlated with times when pain and sedation were minimized.
- P5: Boundary condition: when channels for relief and closure are blocked (e.g. uncontrolled pain or isolation), we do not treat an absence of centering or closure acts as a clean falsifier; the primary falsification test is failure under high-channel

conditions where relief and connection are feasible. Each of these predictions (P1–P5) is framed so it can be tested with real-world data from hospice and palliative settings. And importantly, each has a mirror “red flag” outcome that would undercut LoF:

- RF1: Even with excellent pain control and family support, the final days’ mood scores are systematically worse than expected from prior trend (i.e. people actually decline emotionally instead of centering). If we saw many cases where suffering just deepens at the end despite all support, that would violate the neutral closure idea.
- RF2: We observe *no increase in closure acts* as death nears – even when patients have the chance and support to do so. If, on average, people didn’t make any extra effort to tie up loose ends or connect with loved ones, LoF’s menu-tilt prediction fails.
- RF3: Supposed “counterweight bursts” (lucid rallies, emotional breakthroughs) are no more frequent than similar events in non-terminal phases of illness. For example, if moments of clarity happen just as often in stable patients as in those about to die, then there’s nothing special about end-of-life in this regard – it could just be random.
- RF4: Any positive findings we initially get (say we do see a centering trend or a closure surge) evaporate under closer scrutiny or larger samples. Perhaps what looked like a pattern was just noise or bias – for instance, maybe staff paid more attention to patients who were calmer (a bias in reporting), or maybe a simpler explanation like “people get physiological endorphin surges before death” could account for it without LoF. If a simpler homeostatic or medical model explains the data just as well as our fairness model, then we haven’t proven anything special. We will actively look for these alternative explanations. Additionally, if when we implement blinding and pre-registration the effects disappear (meaning maybe we inadvertently cherry-picked earlier), that’s a strike against LoF.

Observing even one of these red flags consistently would be concerning; observing two or more in well-supported patients would cast serious doubt on the idea of neutral closure. A law must state what would falsify it – and here we have those red lines.

It’s worth noting that modern hospice philosophy already *aligns* with what LoF would predict. Hospice and palliative care focus on widening channels for relief (pain control, emotional support) and relationship (family presence, reconciliation) in the final days. Often, when those channels are opened, patients naturally gravitate toward closure and

find a measure of peace. Our contribution here is to measure that pattern with scientific rigor without altering the care. If LoF is correct, the everyday “small miracles” of hospice – those moments of grace and balance at life’s end – are not just comforting anecdotes but manifestations of an intrinsic balancing mechanism in conscious systems. If LoF is wrong, careful measurement will reveal that too, perhaps by showing that these stories, moving as they are, don’t generalize or don’t exceed what standard biology can explain.

What you’ll get from this Chapter:

- Understanding why the end-of-life is the ultimate test: You’ll see clearly why we consider the final days and hours of consciousness as the make-or-break proving ground for the Law of Fairness. If fairness truly governs life’s outcomes, it *must* appear here in a big way. This chapter explains that intuition and spells out what it means for the theory to “show its hand” as life closes.
- Insight into hospice phenomena through a new lens: We’ll discuss real observations from hospice and end-of-life care—like terminal lucidity (sudden clarity), “last goodbyes,” unexpected moments of calm, or even the well-known pattern of rallying before death—and reinterpret them as potential evidence of a balancing act. You’ll gain an appreciation of how these phenomena, often thought of as mysterious or purely spiritual, could also fit into a scientific pattern predicted by LoF.
- A blueprint for gentle research in a fragile setting: Learn how it’s possible to study psychological patterns at end-of-life *without* ever being cold or intrusive. We describe methodologies that respect every ethical boundary: from ultra-brief mood surveys to passive monitoring, all integrated with standard care. You’ll see how research can be done *with* compassion, not at odds with it, especially in such a sensitive context.
- The exact signs of a “fair” ending (and how to detect them): We enumerate the concrete, measurable signs that a lifetime ledger is balancing out at the end: e.g. emotional neutrality (within a preregistered equivalence band around neutral), spurts of meaningful social interaction, increased focus on comfort and reconciliation, etc. You’ll also see how we quantify those signs (with numbers, rates, and statistical criteria) so it’s not left to subjective interpretation. In other words, we turn the idea of “dying in peace” into something we can actually track and verify across many cases.
- What would prove the idea wrong in this context: Just as crucial, the chapter makes clear what outcomes would invalidate the fairness law when it comes to

end-of-life. This includes scenarios like consistent final despair despite best care, lack of any closure efforts by patients, or patterns that are equally explained by simpler theories (like just the effects of medication). By spelling out these fail conditions, we ensure that the hypothesis is testable and falsifiable. You'll come away knowing exactly what data would shout "No, life doesn't balance at the end" if that's the truth.

- Renewed respect for what remains unknown: Finally, beyond the data and theory, this chapter may give you a renewed perspective on the end of life itself. Whether LoF holds or not, attempting to measure fairness at life's end underscores how remarkable those last moments are. You'll gain an appreciation for how much is *actually happening* in what may seem like quiet, waning days—and how carefully science has to tread in order to learn from them.

Subsections in this Chapter:

- **11.1 Why This Is the Sharpest Test** – Uses intuitive examples and thought experiments to illustrate why a shrinking time horizon makes LoF effects either blatantly obvious or entirely absent. We contrast scenarios of "balanced" versus "unbalanced" final days and discuss how, in theory, a true fairness law would manifest most strongly when there's no time left for error. This section sets up the end-of-life stage as a crucible where only the real deal (or total failure) will show up.
- **11.2 What Hospice Workers See** – Compiles insights and common stories from hospice and palliative care across cultures. It covers recurring themes: patients finding peace after saying goodbye, last visits that seem to uplift patients, terminal lucidity cases, etc. We then tie each of these anecdotes to our predictions (or to rival explanations) to show how they serve as qualitative evidence *and* how they'll guide quantitative measures. Essentially, this section says: "Here's what seasoned caregivers notice in final days, and here's how we plan to capture those patterns in data."
- **11.3 Ethics: What We Will and Will Not Do** – Clearly lays out the ethical framework and specific study design at end-of-life. It assures the reader (and any institutional review board) that our approach honors consent, minimizes intrusion, and prioritizes comfort. We detail things like the consent process, the types of measurements we consider allowable (and those we ruled out as too invasive), and how we'll incorporate feedback from caregivers and families. This section is a transparent guarantee that science will not trump humanity in our project.

- **11.4 Research Notes: Variance Compression and Neural Signatures –** Translates the qualitative predictions into quantitative hypotheses about variability and brain patterns. “Variance compression” refers to the idea that as someone nears death (and if LoF holds), the day-to-day swings in their affect (variance) should shrink compared to earlier in life – we explain how we’d detect that statistically. We also discuss any neural data that could be relevant (for example, if EEG or fMRI are available from patients earlier in illness or from bedside devices, what neural correlates of closing loops or relief might look like). This section might be technical, covering metrics like standard deviation of mood in final week vs. prior weeks, or looking for neural markers of acceptance, but it’s kept as a “research note” so main readers can skip the gory details if they want.
- **11.5 Reading Anecdotes: Scripts vs. Signals –** Offers tools for separating meaningful signals from the human tendency to weave comforting narratives. We acknowledge that end-of-life stories are often told in a hopeful light (the “final moment of clarity” becomes legend, etc.). This section discusses how we avoid being fooled by these scripts. For instance, we might compare reported events against hospital records to check if that “peaceful last day” coincided with heavy sedation (which might explain it without invoking LoF). We introduce the idea of *differentiating correlation from causation* here in anecdotal evidence and emphasize the need for controlled observation. In short, it’s a guide on how not to let wishful thinking color our interpretation of end-of-life phenomena.
- **11.6 Fail Patterns at Terminal Closure –** Details exactly what patterns (or lack thereof) would force us to conclude that LoF’s end-of-life predictions are unsupported. This includes statistical fail criteria like “no significant difference between final-week mood variance and random expectation” or observational ones like “zero instances of reconciliation efforts in X patients who had the opportunity.” We enumerate these failure modes so that, once data comes in, we can promptly recognize if the theory isn’t holding up. This final section is basically our falsification cheat-sheet, ensuring we stay honest about negative results. If the Law of Fairness is going to *show its hand* anywhere, it will be here – and if it doesn’t, we need to be ready to call it.

Where we go next:

We now turn to the evidence itself. In Section 11.1, we start by examining why a rapidly shrinking horizon amplifies the signals (or lack thereof) that LoF would produce. We’ll use simple scenarios to illustrate how, near the end, the difference between a true balancing mechanism and mere wishful thinking becomes stark – setting up the critical tests to come.

11.1 Why This Is the Sharpest Test

When time is long, many narratives can fit the facts. When time is short, far fewer can. End-of-life is where theories about felt balance face their most exacting constraints or else dissolve into vagueness. Even outside the life-span context, the *ending* of an experience heavily influences how we judge its fairness. In a classic experiment, participants undergoing an unpleasant experience (painful cold-water hand immersion) reported the episode as more bearable overall when it lasted slightly longer but ended on a less painful note. The group who had an extra period of milder discomfort at the end rated the whole episode more favorably than those who had a shorter, acutely painful ending (Kahneman, 1993). In other words, a gentler final chapter made the entire ordeal seem “fairer” in hindsight. By extension, the final phase of one’s life can disproportionately color the perceived balance of the whole life: a terminal period of peace and relief may redeem a lifetime of hardship, whereas a bitter end could eclipse decades of otherwise balanced experience. The Law of Fairness makes a crisp, risky claim: as the expected horizon H_t contracts toward the end of conscious life, the internal “menu” of thinkable and permissible actions—formally, the admissible set $\mathcal{A}(t; \bar{L}, H, C)$ —must narrow and tilt toward closure, relief, reconciliation, and meaning-making. If LoF is a real constraint rather than a poetic metaphor, this tilt should become most visible precisely when the horizon is shortest and the system can least afford any move that cannot be compensated.

11.1.1 The mathematics of urgency in plain speech

In Chapters 5 and 6 we introduced the idea of a shadow price λ_t , a multiplier that scales up the value of actions which increase the probability of neutral closure. As the horizon shrinks, λ_t rises. Intuitively: with little time left, every action that reduces “unpaid debts” or unclosed loops (unrepaired relationships, untreated symptoms, unresolved fears) becomes disproportionately valuable, while every action that risks new, unpayable costs becomes prohibitively expensive. Earlier in life, when someone has years ahead, the system can afford indulgent detours because there is ample time to compensate. Near the end, that freedom narrows.

Concretely, under LoF we expect to see three things co-occur when H_t is very small:

Menu shrinkage: Fewer options even *feel* psychologically “alive” or worth considering.

Menu tilt: Within the remaining options, moves that have high “repair” or “relief” value feel easier to initiate and sustain, whereas low-yield detours lose their appeal.

Counterweight concentration: Comfort, lucidity, and reconnection—when the channels for them are open—arrive in brief, intense bursts rather than as diffuse, long phases.

These are not just nice ideas or cultural preferences; they are what a self-regulating system *must* do to keep the lifetime ledger near zero as it approaches the terminal boundary.

Independent lifespan research points the same way. Socioemotional selectivity theory (Carstensen, 1999) proposes that when future time is perceived as limited, goals shift toward emotionally meaningful ends; a robust ‘positivity effect’ has been observed in attention and memory (Mather & Carstensen, 2005). We read these findings as consistent with the horizon-sensitive dynamics under LoF: as H_t contracts and the shadow price λ_t rises, the admissible set tilts toward comfort, reconciliation, and meaning, and both the mean and the variance of $HCI(t)$ should compress toward neutrality after accounting for analgesia/sedation windows.

11.1.2 Why ordinary rival accounts blur here

Common “rival” explanations for behavior—hedonic adaptation, generic physiological homeostasis, trait resilience, etc.—typically predict that people regress toward their own baseline or personality-driven coping style. They *do not* predict a horizon-contingent re-weighting of the choice set. For much of life, especially when someone is healthy with years ahead, these rivals can mimic LoF’s predictions to an extent: after disruptions, mood often drifts back toward an individual baseline, which could be mistaken for “balance.” But at end-of-life, the predictions diverge starkly:

Adaptation/Homeostasis model: With good palliative care, expect the person to stabilize toward their usual mood set-point. Crucially, this model *does not* predict any systematic spike in closure acts or reconciliation attempts as death approaches; it would just expect a return to personal baseline affect if pain is managed.

LoF model: Even with equal symptom relief, expect a selective increase in acts that “close the loops” (e.g. making amends, saying goodbye, tying up loose ends) and a decrease in acts that create new potential debts, with effect sizes growing larger as the horizon contracts.

In short, end-of-life compresses the noise and sharpens the contrast between LoF and its rivals. If LoF is wrong, a dying person’s behavior should *not* systematically change in the repair-weighted way LoF predicts—beyond what adaptation or personality alone would explain. If LoF is right, we’ll see distinctive horizon-driven shifts that rivals do not anticipate.

11.1.3 Natural experiments unique to end-of-life

The end-of-life context provides quasi-experimental contrasts that are unusually informative, often available in routine care:

Predictable vs. uncertain prognosis: Consider two patients with similar symptoms, but one has a clear, short prognosis and the other's timeline is uncertain. LoF predicts a stronger menu tilt (more closure-focused choices) in the patient with the clearer, shorter horizon, even if current mood and pain levels are the same.

Channel opening vs. closing: Consider a sudden improvement in conditions (e.g. effective pain relief administered, or a loved one arriving – channels opening) versus a sudden setback (e.g. strict visitor restrictions imposed – channels closing). LoF predicts immediate shifts in behavior composition: when channels open, the patient should initiate more reconciliations or expressions of gratitude; when channels close, we'd see more quiet, solitary comfort-seeking (and fewer social or relational acts).

Sedation titration cycles: In palliative care, sedation is sometimes carefully adjusted so that patients have alternating periods of wakefulness and comfort. LoF predicts that when a sedated patient becomes lucid (a wakeful window), the *first* actions they take will disproportionately be high-yield acts (e.g. one important communication or closure gesture) rather than arbitrary or trivial actions.

These situations act as built-in A/B tests in care, allowing for within-person comparisons that naturally control for an individual's traits and history. End-of-life thus offers scenarios where we can observe the system under different “horizon” and “channel” conditions in the same person.

11.1.4 What “neutral” means when suffering is real

Saying the ledger should neutralize at the end does not promise the absence of suffering—it's a claim about net balance over the final conscious window. Consider two stylized trajectories:

A life of heavy loss: In the final weeks, suppose the patient's pain is well-controlled, estranged family members have become reachable, and meaningful ritual (spiritual or secular) is available. LoF predicts the person's focus will tilt toward reconciliation attempts, expressions of gratitude, life review, and savoring simple comforts. In quantitative terms, the last days' composite valence (HCl) should end up closer to zero than one would expect from simply extrapolating the prior painful weeks straight to the end. Relief and positive moments disproportionately fill the final window, counterbalancing the earlier pain.

A life of great ease: A person who has had an overwhelmingly positive life might experience gentle “counterweights” near the end—not punishment, but subtle balancing moves. They might show humbling moments of acceptance, make long-neglected apologies, or work to transfer responsibilities and comforts to others, all without causing net harm. The intensity of these counterweights should scale with the degree of imbalance: a very easy life doesn’t require much “downward” correction, but it will still tilt toward closure rather than endless enjoyment.

In both cases, neutrality is an average over the final stretch, not a promise of constant joy or tranquility every second. Agitation, grief, and pain will still occur. The test is whether—given proper access to relief and connection—the integral of experience in the last conscious days centers closer to zero than a naive model would predict. We will allow a small equivalence band around zero (for example, final-week mean within ± 0.15 SD of zero—where SD is the individual’s within-person standard deviation over a pre-terminal reference window—and a final-week mood slope within ± 0.05 SD/day, with variance no more than 80% of prior weeks) as the definition of “close enough to neutral.” If the final ledger falls outside those bounds under high-channel conditions and with adequate data density, LoF fails for that case.

11.1.5 Three measurable signatures of LoF at end-of-life

To keep our claims empirically grounded, we propose three concrete, pre-registrable signatures that would indicate LoF dynamics at work in end-of-life contexts:

Horizon-menu interaction: As independent estimates of remaining time H_t get shorter (e.g. moving to a worse prognosis category or unanimous clinician agreement that death is imminent), within-person logs should show *fewer* novel goal pursuits and *more* closure-weighted acts. In practice, we could track things like the proportion of days on which a patient initiates a new project or venture (expected to drop) versus days on which they initiate a closure act like calling someone or saying goodbye (expected to rise). This is essentially an interaction between horizon and behavior type: shorter horizon \rightarrow a shift toward repair/relief acts.

Burst-like counterweights: In patients who have fluctuating levels of consciousness or clarity, the brief lucid periods should contain a disproportionately high rate of “high-yield” acts (relative to that patient’s own baseline when they were more stable). For example, if a patient becomes lucid for an hour, in that window we might see them accomplish one or two significant acts (an important conversation, a forgiving gesture) that did not occur in similarly long periods before. Statistically, lucid windows would

show an overrepresentation of closure acts compared to random windows from earlier weeks.

Last-window centering: Provided pain is managed and social access is intact (what we'll call "high channel" conditions), the mean HCI during the final 72–96 hours of life will lie closer to zero than what you'd forecast from that person's preceding month. We can formally test this by taking each individual's mood trajectory for the month before their final days, predicting forward to the last few days, and comparing that forecast to what actually happened. LoF predicts the actual final mean will be closer to neutral than the forecast. Moreover, the closer death is (the smaller H_t gets), the stronger this centering effect should become across people.

Each of these signatures can be measured with minimal-burden tools that respect the patient (e.g. very brief mood surveys, simple logs of actions, and extraction of key phrases from clinical notes). Each also comes with clear falsification criteria (outlined later in Section 11.4). They give us specific targets to either observe or *not* observe in data.

11.1.6 Why strict ethics strengthen the test

One might worry that because we refuse any invasive or distress-inducing interventions, our results could be ambiguous ("correlation, not causation"). In fact, the opposite is true: by working only with care-concordant interventions—things that improve comfort and connection, like better analgesia, better communication, more flexible visiting—we eliminate confounds that aggressive research tactics might introduce. If LoF is a real, underlying constraint, we *do not need to engineer suffering to see it*. We only need to widen the channels for relief and observe honestly. A theory that required us to *add* distress in order to prove it would be a poor candidate for a fundamental law of fairness. By demanding that our methods align with humane care, we ensure that any evidence for LoF comes from a natural emergence of balance, not from artifacts of experimental stress.

11.1.7 Boundary conditions clarify the law

If pain is left untreated, if patients are kept in isolation, or if communication is impossible, we expect failures of centering – not because the law ceases to exist, but because the admissible set of actions is artificially starved. These harsh conditions are analytically valuable because they define the limits of LoF: they show us situations where neutrality should not be expected. In other words, the end-of-life window allows us to test both positive predictions (what *should* happen when channels are open and relief is possible) and negative predictions (what *will not* happen when the system's options are cut off).

Seeing neutrality *fail* exactly where we predict it will fail (due to blocked channels) is just as important to confirming LoF as seeing it succeed under optimal conditions.

11.1.8 Why include this chapter in a mainstream book?

Readers deserve falsifiable claims—assertions that could, in principle, be proven wrong. End-of-life offers exactly that opportunity for the Law of Fairness. It compresses time, simplifies choice, and makes the direction of “smart bets” unmistakable. If our predictions do not hold when it matters most—when a person is out of time to adjust—then we should revise or retract the law. Conversely, if these patterns do hold across cultures, care settings, and belief systems (and under rigorous conditions like blinding and preregistration), then we will have seen the Law of Fairness display itself publicly, in a context where no mere comforting metaphor can account for the results.

11.1.9 Where we go next:

From motivation to method. Section 11.2 outlines an observational plan that never interferes with care, defines what counts as a fairness signature, and sets equivalence bounds before any data are seen.

11.2 What Hospice Workers See

The best observers of shrinking horizons are the people who walk those hallways every day. Across countries, faith traditions, and healthcare systems, hospice and palliative care clinicians report a recurrent set of patterns as patients approach their final days. In this section, we translate those field observations into plain-language descriptions and then into testable signals and gentle measurement ideas that correspond directly to the LoF predictions from 11.1.

11.2.1 The “menu tilt” in everyday care

What staff report: As a terminal prognosis becomes clear, many patients grow decisive about a very few things and indifferent to almost everything else. They start to prioritize relief, reconnection, and closure over acquisition, novelty, or long-range plans.

Typical notes in the chart:

“Asked to call brother after 12 years.”

“Wanted quiet music and sunlight; declined TV.”

“Insisted on finishing a brief note to the team.”

LoF interpretation: This is the admissible-set narrowing and tilt in action: as the horizon shrinks, options that provide *relief* or *repair* gain psychological priority and “stickiness,” whereas low-yield detours (like idle distractions or new ventures) lose salience. The patient’s internal menu is focusing on what will close the ledger.

Gentle measures: We could introduce a simple one-page Contact and Closure Log for clinical use, where staff or family note whom the patient actively reaches out to and which “open loops” (unresolved matters) the patient attempts to address each day. Additionally, a brief Repertoire Checklist (a daily yes/no checklist of, say, 10 common acts such as reconcile, bless, bequeath, forgive, savor, pray/reflect, listen to music, sit in sunlight/nature, settle paperwork, *start* a new project) would capture which types of actions a patient engages in. We’d expect, as prognosis shortens, to see more checkmarks in the closure/comfort categories and fewer (eventually none) in the “new project” category.

11.2.2 The last good conversation

What staff report: Patients frequently seem to use a brief period of clarity or energy to say one specific important thing—often an apology, a blessing, final instructions or expressions of love and reassurance like “thank you” or “don’t worry about me.”

Typical phrasing:

“Tell Maria the ring is hers.”

“I forgive you; please forgive me.”

“You’ll be fine. Take care of each other.”

LoF interpretation: This exemplifies counterweight concentration. When time is short and a lucid window opens, the very first things patients choose to say are often those with the highest emotional yield—resolving a key worry, passing on a blessing, or giving permission. In LoF terms: when H_t is near zero, the system selects the highest-yield psychological moves first.

Gentle measures: We could unobtrusively tag the content of such spontaneous statements with one or two keywords (e.g. *apology*, *blessing*, *legacy transfer*, *gratitude*, *reassurance*). We would also record a time stamp relative to key clinical changes, such as the start of effective analgesia or the arrival of a specific visitor, to test whether improved comfort or connection (channel opening) immediately precedes these important utterances. This would let us quantify, for example, if patients are more likely to deliver a “last message” shortly after pain is relieved or a loved one enters the room.

11.2.3 Terminal lucidity and “clear windows”

What staff report: Some patients who have periods of delirium or confusion experience short clear windows—often right after their pain or breathing difficulty is brought under control—in which they become lucid and focused, sometimes surprisingly so given their overall condition. They often use these windows for focused closure activities (for instance, saying something meaningful or making a critical decision).

LoF interpretation: This suggests that when physiological “noise” is quieted (delirium subsides, pain diminishes), the system can once again execute its repair-weighted policies. In other words, once the background chaos is reduced, even a very sick mind will instantiate the same closure-oriented acts predicted by LoF. Terminal lucidity provides an opportunity to see QS in action: with relief in place, even a brief return of clarity is utilized for high-priority acts.

Gentle measures: Introduce a Short Lucidity Scale (e.g. 0–3 rating of alertness and orientation) to be done three times daily by bedside staff. Mark whenever there’s a jump in lucidity (say a 2 or 3 when previous was 0 or 1). Alongside this, include a simple tick-box for whether any targeted act occurred during that window (e.g. *reconciliation call made*, *gave a blessing*, *made a bequest*). With this data, we can do a within-person analysis: does a given patient perform more of these high-yield acts during their clearer

windows than in the periods right before or after? LoF would predict a clear *enrichment* of meaningful acts in those lucid moments.

11.2.4 Permission and timing

What staff report: Some patients seem to “hold on” until a particular person arrives or until they receive explicit permission to let go. Often, a patient will pass away shortly after a long-awaited family member visits or after a loved one says “It’s okay, you can go now.”

LoF interpretation: This could reflect a relational loop that needed closure. A key relationship was part of the ledger, and only when that loop is closed (through presence or explicit permission) does the system allow the final step of letting go. In QS terms, *once the relationship channel is closed properly, the admissible set may now include the act of “letting go” without violating the fairness constraint.*

Gentle measures: Record the arrival times of especially significant people (e.g. a child flying in to see the patient) and note the time of death relative to those events. Also note whether any explicit “permission” or blessing was given (e.g. family saying “*We’ll be okay, you can rest now*”). Then compare cases where such an event happened to matched cases where no last-minute arrival or permission occurred. The test (done without invoking any metaphysics) is whether time-to-death distributions differ significantly. Because family arrivals are often prompted by clinicians’ recognition that death is imminent, analyses should control for clinical indicators of imminence at the time of the arrival/permission event. If patients commonly die soon after a key person’s arrival/permission, more so than those without such events, it suggests that closure of that relational loop might play a role—consistent with LoF’s closure logic.

11.2.5 Dreams, reveries, and the “night workshop”

What staff report: Patients often recount vivid dreams or reveries involving travel, reunions, or symbolic acts of repair. Some wake up with a clear resolution or new understanding—saying things like “I know what I need to do” after a notable dream. Caregivers sometimes notice a patient being more settled or determined following such dreams.

LoF interpretation: As discussed in Chapter 10, sleep and dreaming can serve as a “low-cost counterbalancing” mechanism—a night workshop for emotional processing. Especially when daytime channels for closure are limited (say the patient is too weak for long conversations), the mind may use dreams to work through unfinished business or provide relief imagery. A dream of a reunion could, functionally, be giving the patient a sense of closure with someone who isn’t present. LoF predicts that dreams in terminal

patients will often carry relief or repair motifs and might even prompt daytime actions (like suddenly knowing what needs to be said).

Gentle measures: Offer an optional morning prompt: a caregiver or nurse can softly ask, “Do you remember anything from the night?” If the patient is willing, they can share a short description (one sentence or a few keywords) of any notable dream or thought, possibly into a 10-second audio recorder for accuracy. These snippets can then be minimally coded for *Relief* or *Repair* motifs (for example, “dreamed of traveling to a beautiful place” might be tagged as relief imagery; “saw a departed loved one” might be tagged as a reunion/repair theme). Importantly, we do not interpret or dwell on content with the patient (to avoid suggestion or distress); we simply note if such motifs were present. We can then look at whether these motifs correlate with the patient’s subsequent behavior (did they act on a resolved intention mentioned?) or immediate mood. Consistent with LoF, we’d expect that when daytime avenues are scarce, the “night workshop” picks up some slack in balancing the ledger.

11.2.6 Small pleasures regain gravity

What staff report: Patients in their final days often derive great comfort and joy from the simplest sensory experiences: a certain beloved song, the feel of warm hands holding theirs, a breeze or sunlight on their skin, a sip of a favorite tea. These small pleasures sometimes seem to mean *everything* in that moment.

LoF interpretation: We could call this ReliefGain via low-risk inputs. When horizons shrink, even modest comforts carry significant weight in improving net well-being—because they pose no future cost but provide immediate relief. LoF predicts that as the big ambitions fall away, *small* pleasures should become highly valued (and sought) since they contribute to bringing the affect ledger toward neutral without incurring new debts.

Gentle measures: Provide a simple bedside Comfort Menu listing options like music, gentle touch, fresh air or sunlight, soothing scents, warm blankets or a favorite flavor. Track which items the patient requests or accepts each day, and perhaps have the patient or nurse rate the resulting comfort on a 0–3 ease scale. We’d expect to see *increasing* use of these comfort options per day as H_t falls (fewer days left, more comfort items chosen), and anecdotally, patients might report these experiences as profoundly meaningful or calming. Quantitatively, one could plot requests per day and their comfort ratings: LoF would predict an upward trend in utilization and effective relief from these simple measures as the end approaches.

11.2.7 Reconciliation attempts, successful or not

What staff report: Many patients attempt to reach out to estranged family or friends as the end nears. Not all attempts succeed (some calls go unanswered or past wounds remain), but the frequency of *trying* to reconcile or say farewell markedly rises near the end.

LoF interpretation: The tilting of the admissible set toward repair doesn't guarantee that the other party will reciprocate or that every attempt will succeed. LoF predicts the attempt rate goes up because the person is internally driven to close loops—but the law does not promise that external circumstances will always allow closure. So an increase in *initiated* contacts, even if completion varies, fits the pattern.

Gentle measures: Log each attempted contact versus those that result in actual connection. Note reasons for non-connection (no answer, person cannot be reached in time, relationship refusal, etc.). The key metric from a LoF perspective is the *attempt* frequency. We expect a rising proportion of patients making such attempts as they enter their final days. The success of those attempts may depend on external factors ("channels"), but the act of trying is the signature of LoF. We should see that even when connections fail, the patient at least tried to reconcile more often than they would have weeks or months earlier.

11.2.8 When channels are blocked

What staff report: When key support "channels" are blocked—such as poorly controlled pain, strict isolation (no visitors allowed), bureaucratic obstacles, or even a noisy, chaotic environment—patients tend to become agitated, withdrawn, or appear stuck, and many of the positive patterns (lucid moments, peaceful acceptance) are much rarer. In other words, where comfort and connection are absent, end-of-life often goes poorly.

LoF interpretation: A blocked channel means the system's admissible set is artificially constrained. If pain is overwhelming or the person is completely alone, many compensatory actions can't be executed. LoF would frame these scenarios as the system being prevented from doing its job—hence the ledger may not approach neutral and distress can remain high. These cases are boundary conditions for the end-of-life tests: when channels are blocked (e.g., heavy sedation, isolation, unmanaged symptoms), the chapter's specific signatures are not expected to be reliably measurable, so we treat those periods as non-confirmatory and focus falsification on high-channel windows.

Gentle measures: Keep a simple Barrier Log tracking day by day if any major barriers were present (e.g. pain >6/10 unrelieved, visitor restrictions in place, language barrier without

interpreter, technology failures preventing a key call, etc.). Then compare the “final window” outcomes for high-barrier periods vs. low-barrier periods. We expect centering failures (more negative final HCl, less closure activity) specifically when barriers are high. This is testable as a conditional prediction: LoF says, essentially, *if* Channel Score is high (good access), we should see neutrality; *if not*, we expect it to fail. So observing poorer outcomes with documented high barriers actually *supports* the idea that LoF is a conditional law reliant on available channels.

11.2.9 Pediatric patterns

What staff report: Dying children show some different surface behaviors than adults—often focusing on play, imagination, and comforting rituals—while their parents and families emphasize meaning-making and legacy (handprints, keepsakes, special ceremonies). Yet there’s a sense of closure and peace being sought by everyone involved, just in different ways.

LoF interpretation: The functions are similar, but adjusted for developmental stage. For the child, “repair/relief” may manifest as play (which provides comfort and normalcy) or the need for safety signals and storytelling; for the family, it’s about creating meaning and lasting connections (legacy acts like preserving artwork or making memory handprints). LoF would say the system still strives for balance—relief and connection—but the *form* changes with age-appropriate coping. The child’s admissible set prioritizes feeling safe and loved; the parents’ set includes helping the child’s legacy live on.

Gentle measures: Track something like play episodes for the child (did they engage in play or fantasy or favorite stories today?) and legacy acts by family (did the family initiate any meaning-making activity like a legacy project or community prayer). Test whether these increase as the prognosis becomes more certain or time becomes short. We expect that, as with adults, the frequency of these comfort/meaning behaviors rises when horizons shorten—just tailored to the child’s and family’s context.

11.2.10 Cultural inflections, same structure

What staff report: The *forms* of end-of-life closure differ by culture and spiritual background—some patients pray or want clergy, others sing ancestral songs, others have rituals of forgiveness or story-sharing. Despite the differing forms, the underlying *functions* seem strikingly similar: seeking relief, reconciling relationships, giving blessings, transferring responsibilities or wisdom.

LoF interpretation: We predict cross-cultural invariance at the level of function. That is, while the specific content varies (one culture’s prayer is another’s song; one family values verbal forgiveness, another shows it through action), the categories of relief,

repair, and meaning appear universal. LoF would expect that if we code these acts by their function, we'll see the same rise in Relief/Repair/Meaning acts near the end across cultures. Cultural scripts might shape *how* something is done or described, but not *whether* it's done in service of balancing the ledger.

Gentle measures: For analysis, we code each observed act by its underlying function (Relief, Repair, Meaning, or Neutral), using culture-specific examples for guidance (with input from local advisors to avoid mislabeling). We then test measurement invariance across cultures: ensure that at least configural and metric invariance hold (i.e. the concept of, say, "Relief" behaviors is consistently recognized across groups). If full scalar invariance (equal baselines) isn't met, we restrict comparisons to within-person patterns rather than raw cross-group averages. Only once we establish that our metrics mean the same thing in each culture (Chapter 8 discusses this process) do we compare outcomes. LoF's claim is that *functionally* we'll see the same balancing acts everywhere when conditions allow. If we do all that and still find no increase in closure acts or even opposite patterns in another culture *under equivalent conditions*, that would challenge LoF (see Fail patterns in 11.6.9). But if forms differ while functional trends converge, that supports a universal law operating beneath cultural variation.

11.2.11 The “surge” myth vs. the settling reality

What staff report: Not everyone has a dramatic "last hurrah" of energy or euphoria before death (the popular notion of a big rally). More commonly, staff observe a gentler settling pattern: patients initiate fewer total activities, but the ones they do initiate tend to be the important, high-yield ones. Instead of a broad surge of vitality, it's a focused narrowing with purposeful acts.

LoF interpretation: This aligns with variance compression with priority reweighting rather than a generalized activation. LoF doesn't predict that people suddenly become super-energetic across the board; it predicts that whatever energy or alertness remains will be channeled into a smaller number of more significant actions. Fewer moves, higher average weight per move, and overall a calmer (lower-variance) affective state as extremes are evened out.

Gentle measures: Track the count of patient-initiated actions each day and assign each a simple "closure weight" tag (e.g. +1 for any act that clearly serves Relief, Repair, or Meaning; 0 for neutral acts). As patients enter their final days, we expect the *count* of acts might decline (they're doing fewer things), but the closure-weighted fraction of those acts goes up. In other words, perhaps early in illness a patient did 10 things in a day with 2 being meaningful; later they do 3 things with 2 being meaningful. The raw activity count

dropped, but the proportion of meaningful acts rose—consistent with LoF’s focus on quality over quantity as the end nears. This pattern – lower act entropy, higher average significance – is testable with basic tallies.

11.2.12 Team choreography that widens channels

What staff report: When care teams coordinate to “get everything right” for a patient—meaning pain is well-controlled, the right people are present at the right times, the environment is peaceful, translation or spiritual support is provided as needed, visiting hours are flexibly managed—patients are noticeably more likely to have a final moment of closure or a peaceful death. In essence, when the team actively widens all the channels, the patient often completes one meaningful act (like a final message or gesture) before passing.

LoF interpretation: The Queue System (QS) can only work with the options available to it. A well-choreographed team effectively maximizes the patient’s admissible set by removing external barriers. LoF would predict that in those circumstances, if there is any balancing to be done, the patient is more likely to do it. The team’s role can be seen as setting the stage so that if the LoF-driven process is trying to operate, nothing stands in its way.

Gentle measures: Create a simple Channel Score (0–5) for each day or shift, tallying key factors: pain adequately controlled (yes/no), a key person the patient wants is reachable or present, a quiet private room environment, any needed interpreter/chaplain or cultural support is available, and visiting hours are flexible or needs are accommodated. This gives a score out of 5 for how “open” the channels were that day. We then test whether higher Channel Score days tend to coincide with higher closure activity or better mood centering. LoF would expect a positive correlation: on days where the score is, say, 5/5, the patient is much more likely to perform a closure act or report greater peace. In practice, we might find that nearly all instances of meaningful final acts occurred on high Channel Score days – reinforcing that enabling the system (through good care) is crucial to see the law’s effects.

11.2.13 Vignettes (composite, de-identified)

(To illustrate concretely, consider these simplified composite vignettes, drawn from multiple cases but with identifying details changed. Each demonstrates the pattern of repair then relief, or the role of channels, in line with LoF.)

The call: A man with advanced lung disease, finally comfortable after his breathing difficulty is relieved, asks for his estranged son’s phone number. He gets through and says, “I’m proud of you,” then later rests quietly to his favorite music. Tags: reconciliation,

blessing, sensory comfort. (*LoF: Once relief was provided, he immediately used a clear window to repair a relationship, then sought a source of comfort.*)

The bequest: A retired teacher, declining but lucid, insists on mailing a final handwritten note to a former student. Only after she sees it posted does she relax, requesting soft light and silence for rest. Tags: legacy transfer, closure, comfort. (*LoF: She completed a meaningful act (legacy/meaning) first, then shifted to relief.*)

The blocked channel: A woman in severe pain, with family delayed by travel, grows increasingly agitated and keeps repeating “Not yet, not yet,” refusing to engage or settle. When pain medication finally takes effect and her family arrives, she smiles, whispers “okay,” and soon passes away peacefully. Tags: pain barrier, isolation barrier, closure act upon resolution. (*LoF: While key channels were closed (uncontrolled pain, missing family), she could not find closure. Once those barriers lifted, she immediately achieved her closure (a final acknowledgment) and was able to let go.*)

11.2.14 What to write down (and what not to)

Do capture: objective, time-stamped facts. For example: which acts occurred (and when), which barriers were present, who was with the patient, what comfort measures were chosen, and a brief HCl proxy score. In other words, record the behaviors, events, and scores that we actually measure.

Do not capture: subjective interpretations or grand meanings. Avoid turning observations into narrative judgments like “he needed to make peace” or “a miracle happened.” Do not impose theological or metaphysical conclusions (no “he saw angels” in the data notes, even if that’s part of a patient’s statement—we would instead record *what was said or done* in neutral terms). Also, never pressure a particular narrative (“Did you see a light? Are you at peace now?”). The data should be the acts and experiences themselves, not the story we might be tempted to weave around them. The patient’s or family’s narrative can be honored personally, but in our notes and analysis we code the events.

In summary: The data are the acts, not the interpretations. We should meticulously record what happens and maintain humility about what it means.

11.2.15 How these observations test the law

Each pattern above isn’t just a comforting story – it maps to a quantifiable LoF signature:

Menu tilt: As H_t falls, an increasing proportion of a patient’s acts should fall into the Relief/Repair category (and fewer into trivial/neutral acts). We can measure this as the fraction of daily actions oriented toward closure or comfort.

Counterweight bursts: High-yield acts (reconciliation, blessings, etc.) should cluster in periods of clarity or opportunity (like lucid windows or right after pain relief). Statistically, the rate of such acts per hour should be significantly higher in those windows than outside them.

Last-window centering: The mean HCl in the last few conscious days should be closer to zero (less net negativity or positivity) than what the prior trajectory would have predicted. We'll formally test this via the forecast-versus-actual approach discussed later.

Channel dependence: All these effects should be stronger when barriers are removed (high Channel Score situations). If channels are blocked, we expect not to see menu tilt or centering, which is itself a predicted outcome (a boundary condition of the law).

If these signatures consistently do not appear even under good care (e.g. we find no tilt, no bursts, no centering where we expected them), the claim of end-of-life neutrality is weakened or refuted. Conversely, if these patterns *do* appear across different settings and cultures—and they do so without any coercion or cherry-picking—then the Law of Fairness has revealed another of its “public fingerprints.” In practical terms, we'd have evidence that even in the face of mortality, something systematic is working to balance experiences.

Importantly, all the patterns above can be observed using measures and notes that are already part of good hospice practice or that add minimal burden. Before proceeding to collect any such data, however, we must set out the ethical guardrails. The next section details what we will and will not do in the name of testing a law that is, fundamentally, about fairness and compassion.

11.2.16 Where we go next:

Once the plan is set, Section 11.3 details measures we can ethically gather—mood, reconciliation acts, sleep, pain relief trajectories—and how these inform the ledger without burdening patients.

11.3 Ethics: What We Will and Will Not Do

End-of-life research must embody the very value that the Law of Fairness describes: fairness, especially toward those who are vulnerable. In concrete terms, that means care first, data second—always. This section lays out the commitments and boundaries governing any attempt to observe LoF phenomena in dying patients. We outline the positive duties we have (what we *will* do to protect and respect participants) and the firm prohibitions in place (what we *will not* do, no matter how tempting in the name of science).

11.3.1 First principles

Care primacy: Comfort, dignity, and the established goals of care outrank all scientific aims. If there is ever a conflict between what would yield more data and what keeps a patient comfortable, we sacrifice the data. Research stops or adjusts immediately if it interferes with comfort.

Respect for persons: We obtain informed consent from the patient whenever possible. If the patient lacks decision capacity, we require surrogate consent (from a legal guardian or family) *and* we still seek the patient's *assent*—meaning any affirmative cooperation or non-verbal agreement from the patient. Equally crucial, we employ continuous dissent monitoring: at the first sign that a patient is uncomfortable or wants to stop, even if they can't speak, the procedure ends.

Non-maleficence and proportionality: We only use minimal-burden measures that a reasonable patient would likely tolerate even if no study were happening. This means things like brief questions or scales that take a minute, passive monitoring that doesn't bother the patient, or notes from routine care. If a procedure would cause more than minimal discomfort or annoyance to a dying person, it's off the table.

Justice: No group of patients should bear disproportionate burden or risk for the research. We also ensure equitable access to any potential benefits of the research: for example, if a comfort intervention is introduced as part of a study, all similar patients should have access to it, not just “research participants.” We avoid any hint of exploiting one population for knowledge that only benefits others.

Epistemic humility: We approach each situation knowing we could be wrong. We measure observable acts and experiences; we do not presume or impose grand meanings on them. We remain open to being surprised or disproven. In practical terms, this principle reminds us not to treat patients as vehicles to confirm our theory—our job is to observe carefully and let the data speak, even if it contradicts our expectations.

11.3.2 What we will do (allowed, with safeguards)

Opt-in micro-measures: We will use only very short measurement tasks that patients explicitly agree to. For example, a simplified HCI mood check (perhaps four questions taking \leq 60 seconds) once in a while when the patient is awake and willing. Or a one-page checklist for contact/closure acts that can be filled out by family or staff. Or letting the patient pick comfort measures from a menu (which doubles as data on their choices). Every measure is optional, brief, and designed to feel like part of care rather than a separate exam.

Passive chart abstraction: We make full use of data already being recorded in medical charts—symptom ratings, medication times and doses, notes about visitors, etc.—by extracting and aggregating it (with proper privacy safeguards). This yields a lot of information without asking the patient to do anything extra. All such data will be de-identified in analysis.

Family/caregiver proxies: If patients cannot self-report (too fatigued or non-communicative), we may invite a willing family member or caregiver to provide simple ratings on the patient’s behalf (e.g. “On a scale from 0–10, how at ease did they seem today?”). We will only do this if it doesn’t create burden or conflict for the family. The surrogate’s perspective can add insight, but we won’t pressure them for data.

Optional narrative snippets: If patients spontaneously share a dream or a meaningful thought, with their permission we’ll record a brief note or audio snippet of it. We’ll then tag it for analysis (as described earlier) but *without interpreting it* to them. We will not push them to talk—only capturing what they *freely* choose to express.

Channel-widening supports: We actively implement tools or steps that not only help the study but directly benefit the patient’s care. For example, coordinating schedules so family calls can happen (and keeping a log of those calls), placing “quiet hours” signs to ensure a peaceful environment (and noting noise levels), or providing translation services for non-native speakers to improve communication. These interventions improve care and incidentally provide standardized conditions that are good for data. Essentially, wherever we can, we turn research needs into care improvements.

Preregistration and blinding: All our analysis plans will be preregistered in advance, and whenever possible, those analyzing the data or coding qualitative information will be blinded to key outcomes. For instance, someone coding the content of patient quotes won’t know whether that patient died in distress or peace (to prevent bias in interpretation). We’ll also pre-specify which analyses we’ll run, so we’re not fishing around after the fact.

Return benign value: We will offer, to any patient or family who wants it, a simple summary of the positive things we recorded. For example, if the data shows a timeline of meaningful moments (“made amends with brother Monday; enjoyed music Tuesday; said he was at peace Wednesday”), we can compile that into a little keepsake or summary letter after the fact. This would be done only with permission and with sensitivity, and only if it’s something the family finds comforting. It ensures that participants potentially get something of personal value from the study (besides altruism).

11.3.3 What we will not do (prohibited)

No symptom withholding or manipulation: We will never delay, reduce, or interfere with indicated treatments (pain relief, anxiety meds, oxygen, sedation, etc.) just to see what happens without them. If anything, our research stance is to enhance symptom control (since LoF predicts better outcomes with better symptom control). Under no circumstance would we hold off on analgesia or adjust a dose for data-gathering reasons.

No coercion, pressure, or scripting expectations: We will not use language that suggests patients “should” be doing certain things (no “This study is about finding peace—do you feel you have to forgive someone?”). We avoid anything that imposes an expectation of reconciliation or a “good death.” Participation must never make a patient feel judged or obligated to act out a narrative. We explicitly instruct staff not to prime patients with ideas like “many people have a last burst of energy” or any teleological framing. We measure what happens; we don’t script it.

No waking patients for research: If a patient is sleeping or finally comfortable, we will not wake them up to ask questions or run a test. Sleep and rest are part of care. Research activities will only occur when patients are already awake and at relative ease. Even then, if a patient looks tired or disengaged, we postpone or cancel the measurement.

No invasive add-ons: We will not introduce any invasive procedure that isn’t part of normal care just for research. That means no extra blood draws, no spinal taps or additional IVs, no experimental device attachments beyond perhaps a soft wearable if the patient desires it. Certainly no neurostimulation or heavy monitoring that could cause discomfort. If something isn’t already benign enough to be routine in hospice (like maybe a pulse oximeter or a fitness-band-like monitor if the patient is curious), we won’t impose it.

No metaphysical adjudication: We will not treat any patient’s subjective experiences (visions, final words, etc.) as proof of metaphysical claims, nor will we try to convince

them otherwise. For example, if a patient says “I saw my late husband in a dream,” we don’t label that in the data as “visitation from beyond” or conversely as “hallucination.” We simply note what was said/done in neutral terms (e.g. “spoke of seeing husband in dream, appeared calm after”). We’re not in the business of affirming or denying spiritual interpretations in the context of care—we leave those judgments out of the dataset.

No post-hoc consent sleight-of-hand: If a patient was confused or delirious and something was recorded during that state (with prior consent), but later the patient (or surrogate) would not have wanted that included, we exclude it. We won’t use data from periods of incapacity unless it was within a clearly consented plan and does not violate any participant’s trust. We never operate on a “forgiveness rather than permission” ethic—especially here. If clarity returns and a patient withdraws consent or expresses discomfort about what’s been recorded, we honor that and remove their data.

No publication of identifying details: When we publish case studies or vignettes, they will be composites or fully anonymized. We strip out dates, locations, unique family details, or rare disease identifiers that could inadvertently reveal someone’s identity. If there’s ever a desire to quote a patient or family member verbatim (something particularly insightful), we will only do so with explicit permission and after ensuring it cannot be traced back to them. The default, however, is to present aggregated or illustrative cases that don’t correspond to any one individual in full detail.

11.3.4 Consent, assent, and ongoing permission

Tiered consent: Our consent forms (for patient or surrogate) will be modular. Each type of data collection is a separate checkbox or line item. For example, one checkbox for brief surveys, one for using wearable data, one for audio-recording any spoken reflections, etc. Participants can agree to some but not others. Importantly, saying “no” to any one part *never* disqualifies someone from being in the study for the other parts. This respects individual comfort levels and ensures no one feels they have to accept a disliked procedure just to participate at all.

Assent/dissent training: All staff involved are trained to recognize the subtle signs of assent and dissent in patients who may not speak. Assent might be indicated by things like a patient maintaining eye contact, nodding when asked about participation, or willingly answering questions day after day. Dissent could be a grimace, turning the head away, closing eyes, agitation when approached with a question—any sign of “I don’t want this now.” Staff are instructed that dissent ends the procedure immediately, no matter what the surrogate consent might say. Ongoing comfort is key.

Surrogate roles: When the patient lacks capacity and a surrogate (usually next of kin or legally authorized rep) gives consent, we treat that consent as *necessary but not sufficient*. We continually cross-check with whatever we can gauge of the patient's own will (body language, etc.). Surrogates will be briefed frequently and asked to renew or confirm consent at intervals (since situations change quickly in end-of-life). We also make sure surrogates know they can change their mind at any time too. Flexibility is maintained—if a family feels something is no longer in line with the patient's values, we adjust or stop.

Right to withdraw: Any patient or surrogate can withdraw from the study at any point without needing to explain why. If withdrawal happens, we stop all data collection immediately. We will also remove or destroy any data that was collected but not yet analyzed, unless they permit us to use what was gathered up to that point. Essentially, participation is entirely voluntary up to the very end, and opting out has no penalty or prejudice on the care they receive.

11.3.5 Cultural and spiritual humility

Function over form: As mentioned, we code and analyze behaviors by their function (relief vs. repair vs. meaning), not by the specific cultural form they take. We remain careful not to privilege one form over another. Praying with a priest, singing with family, performing a tribal ritual, or simply talking through life events—all might serve a meaning-making function. Our job is to respect each form and translate it into our functional coding without bias. (E.g. both praying and singing could be coded as “meaning” if they serve that role for the patient.)

Local advisors: Every site or community we work with should have a Community Advisory Panel or at least consultants that include people like a hospice chaplain, a cultural liaison familiar with local customs, maybe a patient advocate or a family member from the community. This panel reviews our instruments and language for cultural sensitivity and relevance. They help identify if any question or measure might be inappropriate or misunderstood in that context, and they suggest better approaches. Essentially, we don't parachute a one-size-fits-all tool; we adapt with local guidance.

Language care: We avoid judgmental or directive language in both our interaction and documentation. For instance, we would not write “Patient refused to reconcile with X” (too value-laden); we'd write “Patient declined a visit/phone call with X.” Also, we avoid “ought” statements in conversation—no “you should” or “you need to” from researchers. We use patient-centric descriptions: “Patient asked to call her sister” rather than

“Patient sought forgiveness from sister” (the latter assumes too much). By being careful with wording, we reduce the chance of imposing our own framework on their experience.

11.3.6 Data protection and stewardship

De-identification pipeline: All data (notes, logs, recordings, etc.) will be stripped of direct identifiers as early as possible. We assign random study codes; real names are kept separate in a secure key. Audio recordings will have any personal names or places bleeped in transcripts. Dates and exact times might be blurred (e.g. just “Day 3 of study” rather than July 5, 2025). Only the minimal necessary clinical info (like “patient had metastatic cancer” or an age range) is kept with the dataset. The mapping from code to identity is kept on a separate, encrypted system with very limited access, just in case we need to link back for data withdrawal or clinical correlation.

Access tiers: We set up multiple levels of data access rights. Raw data (especially anything with potential identifiers, like full transcripts or videos if any) is locked down to the core research team on a need-to-know basis. De-identified, processed datasets can be accessed by a broader set of investigators under IRB (ethics board) approval. Finally, only aggregate results or completely anonymized summaries leave the team for publication. If we ever share data publicly (for transparency or open science), it will be at the summary statistics or heavily anonymized level, unless participants explicitly gave permission for more (which is rare in this context).

Audit trails: We maintain an immutable log of who accesses the data and any changes made to data (for example, any cleaning or exclusion). Regular external audits (perhaps by an independent data monitor) will verify that our procedures match what we said we’d do. Given the sensitivity, we want a documented chain for every piece of data from collection to analysis. This is part of ensuring trust—so that if questions ever arise (“did you exclude certain patients’ data unfairly?” or “who listened to this recording?”), we have a clear record.

Data minimization and expiry: We only collect what we genuinely need to test the hypotheses (and what participants have agreed to). We’re not fishing for all possible data. If something isn’t directly related to our outcomes or covariates of interest, we likely won’t record it. We also set default data retention policies: for example, identifiable data might be scheduled for deletion after a certain period post-study (unless consent was given to retain for future research). This prevents indefinite holding of sensitive information. Participants might also be given choices about this: e.g. “Would you like your contribution to be destroyed after the study or saved (de-identified) for future analysis of related questions?” honoring their preferences.

11.3.7 Adverse events, triggers, and stop rules

Burden triggers: We will continuously monitor whether participation is causing any distress or burden. For instance, if we notice that patients who participate have higher agitation scores or report feeling overwhelmed on days with study tasks, that's a red flag. Or if staff report that collecting data is delaying care (even by minutes), that's not acceptable. We will set specific criteria (like, if a patient's distress score increases by more than X on days after measurements, or if >Y% of nurses feel the study is interfering) to trigger a pause and review of the protocol. The welfare of participants comes first; if our methods inadvertently add burden, we redesign or halt them.

Environment triggers: External circumstances can also force a pause. For example, if a ward goes into COVID lockdown or there's a staffing crisis, or maybe a patient's room becomes disruptive due to hospital noise or emergency events elsewhere, we would suspend research activities during that period. Another example: if visitor bans are instituted (like in early pandemic) and that obviously conflicts with our idea of keeping channels open, we wouldn't push through with normal protocol as if nothing changed. We'd either adapt or pause, because the study conditions have deviated from what we consider ethically acceptable (no family allowed = we aren't going to test LoF in what we consider a hamstrung scenario without recalibrating).

Futility/harms stop rule: We will predefine interim analyses points (say after every 20 patients or periodically in a long-running study) to check two things: (1) that our interventions/measures are not harming care, and (2) whether the data is showing any potential trend. If after a reasonable number of patients the results are essentially null and there's no improvement in sight with larger N (for example, effect sizes are near zero and confidence intervals tight), we might stop early for futility – no point subjecting more patients to participation if it looks like we won't get a meaningful result. Conversely, if something unexpectedly is making care worse (maybe even the tiny burden we add is causing measurable stress), we'd stop on ethical grounds. In short, two consecutive interim reviews showing “no benefit to knowledge *and* some burden” would terminate the study at that site. And any clear harm signal stops things immediately.

11.3.8 Dual-use and communication ethics

No policy weaponization: We commit that any findings from this research will never be misused to *restrict* patient care under the guise of “improving fairness.” For example, if we find evidence for LoF patterns, no one should twist that into “let's withhold morphine to see if they reconcile” or any ghastly policy like that. In all reports and recommendations, we explicitly state that results *must not* be used to justify reducing

analgesia, limiting visitation, or prolonging life against a patient's will "to achieve neutrality." In fact, our messaging will underscore the opposite: that maximizing comfort and dignity is part of the natural balancing process. We will actively lobby against any misinterpretation that could harm patients.

Care-forward messaging: When communicating to the public (or even to participants and families), we frame everything in terms of improving care and understanding patients, *not* in terms of some test that patients either pass or fail. We never say anything like "we're testing whether this person dies with a balanced ledger." Instead, we emphasize that we're observing how care practices (like good hospice techniques) affect patient well-being. We highlight that the goal is to ensure everyone has every chance for comfort and closure, not to judge anyone's end-of-life. Essentially, any press releases or talks will talk about hospice best practices and patient dignity, not "proving our theory right." If LoF is real, it will show through good care, not through framing someone's death as an experiment.

Publish nulls: We commit in advance to publishing negative or inconclusive results with the same transparency as positive ones. If we do all this work and find no evidence of LoF patterns or even contradictory evidence, we will report that openly—along with all the context needed. We won't bury a "failed" study just because it doesn't confirm the theory. This is critical both ethically (participants deserve to know that their contribution wasn't hidden away) and scientifically (to avoid publication bias and let others learn from what we did). Our preregistrations will include a vow to publish or publicly post results regardless of outcome.

11.3.9 Staff well-being

Debriefing and rotation: We recognize that staff (nurses, researchers, etc.) involved in end-of-life studies can experience emotional and moral strain. Dealing with dying patients and balancing roles can take a toll. We will schedule regular debriefing sessions—safe, confidential meetings where staff can share feelings, discuss any distress, and support each other. We also plan to rotate duties so that no single staff member is continuously tasked with, say, conducting sensitive interviews or being present at many deaths in a short time. Rotation ensures people have time away from the intensity to recover.

Training: All team members will undergo brief but focused training on topics like how to handle assent/dissent, how to be culturally sensitive, and how to listen non-directively (letting patients lead conversations, not injecting our own agenda). Role-playing scenarios will be used, e.g. practicing how to respond if a patient asks "Why are you

studying this? Do you think I'm dying?" or if a family member is anxious about the research. We also train staff to be vigilant for their own burnout signs and to speak up if they need a break or help. Ensuring the well-being of the caregivers and researchers is part of the ethical conduct of the study—burnt-out staff cannot provide the compassion and care required.

11.3.10 Governance

Independent oversight: We will have an Independent Review Board (IRB or equivalent ethics board) approval as a baseline, but beyond that we'll set up a Data and Dignity Monitoring Board (DDMB). This board would include at least one hospice/palliative clinician not otherwise involved in the study, one ethicist, one patient family advocate (someone who has had a loved one in hospice, for instance), and one methodology expert. This board has the authority to review data periodically and demand changes or stop the study if ethical or data-integrity issues arise. They'll look at both safety (dignity preserved? any signals of distress?) and scientific rigor. Their independent eyes make sure we don't become blind to problems because of our involvement.

Open preregistration registry: Every study protocol will be logged in an open registry accessible to other researchers and the public. This includes a plain-language summary of what we're doing. We'll update this registry with any changes or important events (all timestamped) so there is an external record. This transparency helps build trust and also forces us to stick to what we said we'd do (or openly admit if we change something).

Adversarial collaboration invitation: We will actively invite rival frameworks (like proponents of pure "it's all just medication effects" or other psychological theories) to weigh in and even pre-specify analyses that would test their perspective against ours. And we're committed to sharing the de-identified data (under proper agreements) with those independent analysts for head-to-head comparisons. This way, if someone thinks they can explain the data better with a different model, they have the opportunity. For example, we could have one team test a pure "hedonic adaptation" model and another test the LoF model on the same data and compare results under neutral conditions set by a third party. All of this would happen under strict governance to protect participant data, but the point is to avoid siloed, biased analysis. We want the truth, not just confirmation.

11.3.11 Why these guardrails strengthen the science

A theory about fairness and balance must be testable without committing unfair or harmful acts in the testing. By holding ourselves to these ethical guardrails—care-concordant, patient-centered observations only—we actually improve the science.

Here's why: we drastically reduce confounds that might come from research intrusion (because we aren't intruding much). We *increase external validity* because any patterns we detect are arising under real-world compassionate care, not an artificial lab setup. And any positive result we get is more believable: if LoF's signatures appear even when we've been this careful, skeptics will have a hard time dismissing it as an artifact of coercive or biasing methods.

On the flip side, if we do not see LoF effects under these gentle conditions, the correct response is *not* to tighten the screws on patients to "force" a result—that would violate everything. Instead, it would mean the theory might simply be wrong or incomplete in this domain. We would have to revise the theory rather than our ethics.

In short, designing a study that *could* fail while still treating patients humanely is the only acceptable path. It ensures that if evidence for LoF emerges, it is genuine and not purchased at the cost of human suffering. And if the evidence doesn't emerge, we have protected those in our care and can walk away knowing we did no harm in finding out. Either outcome yields insight: a confirmed pattern under humane conditions would powerfully support LoF, whereas a null result would prompt us to rethink the law's claims about end-of-life. Both advances are a public good, achieved without ever betraying the dignity of those we aim to learn from.

Bridge to 11.4: With ethical guardrails firmly in place, we can now talk about the specific signatures we expect at end-of-life—like reduced variability in actions ("compression") and brain/physiological patterns associated with these high-priority acts—and how we'd detect them without adding burden. The next section delves into the Research Notes: turning the qualitative observations and theoretical claims into concrete metrics, models, and neural correlates we can analyze.

11.3.12 Where we go next:

Design alone isn't enough. Section 11.4 covers analysis choices, preregistration, and blinding—how we keep hopeful stories from outrunning what the data can truly say.

11.4 Research Notes: Variance Compression and Neural Signatures

This section translates the end-of-life predictions from 11.1–11.3 into concrete quantitative hypotheses: what metrics to compute, what analytic models to use, and which neural/physiological signatures to look for, all using low-burden, hospice-compatible methods.

To formalize our approach, recall how we define the *ledger*:

$$\hat{L}(t) = \int_0^t HCl(\tau) d\tau$$

the cumulative estimated hedonic balance up to time t (with HCl as our proxy for momentary net affect). The theoretical lifetime ledger (true but unobservable) is

$$L(T) = \int_0^T F(t) dt$$

the integral of the actual net affect $F(t)$ over the stream of consciousness from start to end. LoF's neutrality claim implies $L(T)$ should be approximately zero for an admissible life. In practice, we work with $\hat{L}(t)$ as the measurable surrogate and ask whether $\hat{L}(T)$ is near zero at the terminal time. With those definitions in mind, we now specify testable hypotheses A and B for behavioral data, and then additional hypotheses for neural patterns and other signals.

11.4.1 Hypothesis A — variance compression of behavior

Claim: As the horizon H_t shrinks, a person's behavioral repertoire narrows (fewer distinct types of actions) and their priorities shift toward high-yield relief/repair acts. This is not just global apathy or “giving up,” but a focused *funneling* of effort. We expect to see reduced variety/entropy in daily activities (many things drop out) *and* a reweighting toward meaningful acts among those that remain.

Minimal observables (collected daily, opt-in):

Act stream: A time-stamped log of a small, standardized set of possible patient-initiated acts. We can predefine 10–12 categories that cover common end-of-life actions (e.g. reconcile call, give blessing, bequest or legacy act, listen to music, sit in light or nature, accept touch, pray/reflect, handle paperwork, start a new project, etc.). Each day, we mark which of these occurred. This gives a simple record of *which* types of acts the patient did that day.

Closure tags: For each act in the log, we label it by function: Relief, Repair, Meaning, or Neutral (as defined in Section 11.5.7's vocabulary). For example, “listened to music” gets Relief; “called estranged brother” gets Repair; “wrote a letter to grandchildren” gets Meaning; “watched random TV” would be Neutral.

Hedonic proxy: A brief 4-item HCI check-in (as described earlier: pain, breath ease, emotional ease, connectedness, each 0–10) to track daily net affect. This could be self-reported or via caregiver proxy if needed.

Channel score: The daily Channel Score (0–5) as defined in 11.2.12, indicating how open channels were (pain controlled, key person present, quiet environment, etc.). This is a crucial covariate to ensure we interpret results under “good care” conditions.

Horizon band: A categorical estimate of how close to death the patient is, updated as needed. This could be based on clinician consensus or prognosis categories (e.g. “weeks to months” vs. “days to a week” vs. “hours to days”). It’s an independent variable approximating H_t .

From these, we derive metrics of interest:

Repertoire size R_t : The count of distinct act categories the patient initiated on day t . (E.g. if they only listened to music and talked to family, $R_t = 2$ for that day.)

Act entropy $H_{\text{acts},t}$: The entropy of the distribution of act types on day t . (Here $H_{\text{acts},t}$ is Shannon entropy and should not be confused with the horizon H_t .) This is calculated as $H_{\text{acts},t} = -\sum_i p_{\{i,t\}} \log p_{\{i,t\}}$, where $p_{\{i,t\}}$ is the proportion of the day’s acts that fell into category i . If a patient did 5 acts and 4 of them were the same type, entropy is low; if all were different types, entropy is higher. Entropy captures both repertoire size and evenness of usage.

Closure weight W_t : A weighted sum of acts that gives +1 for each Relief, +1 for Repair, +1 for Meaning act, and 0 for Neutral acts on day t . Formally, $W_t = \sum_i w_i a_{\{i,t\}}$, where $w_i = +1$ if act i is of type Repair/Relief/Meaning (i.e., contributes to closure) and $w_i = 0$ if Neutral, and $a_{\{i,t\}}$ is the count of times act i occurred on day t (0 if none). In effect, W_t is just the count of “closure-relevant” acts that day.

Yield-per-act Y_t : The change in HCI (net affect) per act that day. More specifically, one could compute $\Delta HCl_t = HCl_{\text{end_of_day}} - HCl_{\text{start_of_day}}$ (or compare to a baseline), then divide by the number of acts that day: $Y_t = \Delta HCl_t / \# \text{acts}_t$ when $\# \text{acts}_t > 0$ (otherwise treat Y_t as missing). This is a rough measure of how much affect was improved or stabilized per action taken. A high Y_t could mean fewer acts but each had a big positive impact on mood (or prevented a decline).

Predicted pattern (qualitatively):

As the remaining horizon H_t decreases (i.e. moving from “months” to “weeks” to “days” to “hours”):

R_t (daily repertoire size) should decrease – patients engage in fewer categories of activity.

$H_{acts,t}$ (entropy of acts) should decrease – their day's activities become less diverse/more concentrated. (*This is the “variance compression” in behavior.*)

Meanwhile, W_t (closure-weighted acts) should increase – even if total acts drop, the count of meaningful acts doesn't drop as much, possibly even rises proportionally.

Y_t (affective yield per act) should increase – if they do less but focus on what helps, the emotional benefit (or maintained balance) per action is higher. This is the “priority reweighting” aspect.

Crucially, all these trends are expected conditional on Channel score being high (i.e. under good care conditions). If Channel score is low (barriers present), these patterns might not hold – which is itself informative (LoF would say the system was hamstrung).

Model (preregistered approach):

For compression outcomes like act entropy or repertoire count, we can use a linear mixed-effects model. For example:

$$H_{acts,t} \sim \alpha + \beta_1 H_t + \beta_2 \text{Channel}_t + \beta_3 \text{HCl}_{t-1} + u_{\text{person}} + \varepsilon_t.$$

This model predicts daily act entropy from the current horizon estimate H_t (coded as remaining time, so larger H_t = more time left), the Channel score that day, and the prior day's HCl (to account for mood influencing activity level). We include a random intercept u_{person} for each person to handle individual differences, and residual ε_t .

Under the compression hypothesis, shorter horizons (smaller H_t) correspond to lower entropy. Therefore, when H_t is coded as remaining time, we expect $\beta_1 > 0$: larger H_t (more time left) \rightarrow higher entropy; equivalently, as $H_t \downarrow$, $H_{acts,t} \downarrow$. A significant $\beta_1 > 0$ would support the compression hypothesis.

If instead we model urgency explicitly using $H^{-1}(t)$ (the reciprocal of remaining time), then we expect the coefficient on $H^{-1}(t)$ to be negative for entropy, since greater urgency (larger $H^{-1}(t)$) should correspond to lower entropy. In that parameterization, a significant negative coefficient would support compression. (In practice, we may use $H^{-1}(t)$, log-transformed time remaining, or categorical horizon bands depending on distributional fit; the directional prediction remains the same.)

For reweighting outcomes like W_t , we expect a non-linear effect of horizon because urgency accelerates as remaining time shrinks. Using $H^{-1}(t)$ to model this acceleration:

$$W_t \sim \alpha + \gamma_1 H^{-1}(t) + \gamma_2 \text{Channel}_t + \gamma_3 \text{HCl}_{t-1} + u_{\text{person}} + \varepsilon_t.$$

Here we expect $\gamma_1 > 0$: as the horizon shortens ($H^{-1}(t)$ increases), the number of closure-weighted acts W_t increases. This reflects priority reweighting under shrinking time horizons.

We might also include a Horizon \times Channel interaction term: we anticipate the effect of horizon is strongest when channels are open (i.e., under good conditions, people strongly increase closure acts; under bad conditions, the effect may be blunted).

An interaction term $\text{Horizon}^{-1} \times \text{Channel} > 0$ would support that.

(Technical note: Since R_t is a count of distinct acts, one could model it with Poisson regression. If dispersion > 1.2 , we'd switch to a negative binomial with a log link. For completeness, we will cross-check models of R_t specifically using count regression to confirm they align with entropy results.)

Smallest effect sizes of interest (SESOI):

We will define ahead of time what magnitude of effect would be considered meaningful. For example, for entropy compression: we might set a SESOI that the entropy in the *shortest horizon band* is at least 0.25 SD lower than in the longest horizon band (this is a medium-small effect). For reweighting: maybe that W_t is higher by the equivalent of ~ 0.4 extra meaningful acts per day in the last week of life compared to a month earlier (after covariates). These thresholds will guide our equivalence tests and power calculations.

Falsifier for Hypothesis A: If we find no reduction in repertoire/entropy and no increase in closure weighting even when *Channel scores are high*, that contradicts LoF. Specifically, if our mixed models yield $\beta_1 \leq 0$ (horizon not reducing entropy) and $\gamma_1 \leq 0$ (horizon not boosting closure acts) in two independent cohorts and with confidence intervals excluding our SESOI, then the prediction fails. Another failure mode: if we do see fewer acts as time shrinks but it's accompanied by a drop in closure weighting as well (i.e. patients just do less of everything, consistent with a fatigue/apathy model). If, for example, act counts drop but W_t does not rise or even falls, then what we're seeing might just be physical decline, not LoF-driven prioritization. A strong falsification would be evidence that any behavioral compression is explained entirely by general fatigue or depression (e.g. counts drop and the proportion of meaningful acts drops or stays flat). In statistical terms, a rival "fatigue/apathy" model that predicts lower act count but no change in composition would fit as well as or better than our LoF model.

11.4.2 Hypothesis B — last-window centering

Claim: Given open channels (pain managed, etc.), the mean valence of the final conscious 72–96 hours will be closer to neutral (zero) than one would expect based on

that person's prior trajectory. In other words, in the last days there is a leveling out toward an emotional balance.

Operational test design: We use each patient as their own control by building a personalized forecast of their end-of-life mood based on earlier data, then see if the actual outcome deviates in the direction of neutral. Concretely:

State-space or time-series forecast: Take the time series of each patient's daily HCl (or another affect measure) up to, say, 4 days before death. Fit an individual-specific model—this could be a simple autoregressive model, a Kalman filter, or any reasonable forecasting method (even a clinician's prognosis of mood trend, but we'll likely use ARIMA or Kalman). This model gives a predicted distribution for the final 3–4 days' average affect and a 95% prediction interval (PI).

Compute outcome metrics:

Let \bar{HCl}_{last} be the observed mean HCl over the last 72–96 hours of consciousness (here \bar{H} denotes an average over that window).

Let 0 represent neutral affect. We define:

$D_0 = |\bar{HCl}_{last} - 0|$, the absolute distance from neutral in the final window.

$D_f = |\hat{HCl}_{forecast} - 0|$, the forecasted absolute distance from neutral for the final window (based on the pre-terminal trajectory). Prediction: For most patients with high channel scores, $D_0 < D_f$. That is, the observed final mean is closer to neutral than the forecasted final mean. We also expect, on a group level, that the average D_0 will be smaller than the average D_f . In more statistical terms, if we examine $D_0 - D_f$ for each patient, we expect this value to be generally negative and significantly so when aggregated. A paired test across patients comparing D_0 vs. D_f could be used. We define a “medium” effect as a clearly appreciable difference — for example, at least a 0.3–0.5 SD reduction in favor of D_0 being smaller.

In plainer language: the actual final days should not be as extreme (in emotion) as one would think based on how things were going before—they should be more tempered (closer to okay/neutral).

Falsifier: If under high Channel conditions we find that D_0 is not typically smaller than D_f —for instance, if many patients' final affect is just as far or farther from neutral as their prior trend predicted—then centering did not occur. The clearest failure would be if we often see $D_0 \geq D_f$ (no reduction in distance, maybe even more divergence). Particularly, if we find in a well-powered sample that the paired difference $D_0 - D_f$ is around zero or positive on average (and our equivalence tests rule out more than a trivial negative

difference), that undermines LoF's neutral closure claim. Another red flag is if any centering effect disappears under stricter analysis: e.g., it might seem like centering until we account for a particular variable or remove a bias, and then it vanishes (or if initial positive results don't replicate in a confirmatory dataset).

We'd also incorporate the condition that this test is only valid "under high Channel scores." If some patients didn't have good symptom control, they might legitimately not center (which LoF actually allows). So we would likely analyze a subset of patients meeting care quality thresholds for this test. Failure in that subset is what counts strongly against LoF.

11.4.3 Burst-like counterweights in clear windows

Claim: In patients with fluctuating mental status, lucid windows (periods of clarity) are enriched with high-yield acts compared to adjacent periods of confusion or unresponsiveness.

We operationalize this with a within-person comparison:

Use the Lucidity Scale data (0–3 thrice daily, as described earlier). Mark each time period as a "lucid window" if lucidity rating is high (e.g. 2 or 3) and basic comfort is present (pain controlled, etc.). Adjacent lower-lucidity periods serve as a control.

For each patient, calculate the odds ratio of a high-yield act (Repair/Relief/Meaning act) occurring in a lucid window versus in a non-lucid period.

Prediction: The odds ratio should be substantially > 1 . For example, we might expect OR > 1.8 or so, meaning the patient is nearly twice (or more) as likely to do something meaningful in a clear period than otherwise. In aggregate, across patients, we'd test whether this OR is significantly above 1.

We also account for factors like family presence – since a lucid period when family are present naturally has more opportunity for a meaningful interaction. So we might refine: compare windows to non-windows *controlling for whether a key person was present*. Ideally, we'd use a logistic mixed model for each patient's data: outcome = whether a meaningful act occurred in that interval, predictor = lucidity (yes/no) plus maybe indicator of visitor presence, with random effect for patient.

Falsifier: If the odds ratio is ~ 1.0 (no enrichment) across persons, i.e. patients are no more likely to do important acts during clarity than at random, that contradicts the idea that clarity is used for counterweights. If analysis shows any observed enrichment was entirely due to family presence (for instance, maybe lucidity and family arrival often coincide, and it's really only the arrival that matters), and a model attributing it to lucidity

has no effect once that's accounted for, that weakens the claim. Essentially, if lucidity itself doesn't carry an independent association with meaningful acts, LoF's notion of "the system jumps on clear moments" is not supported.

11.4.4 Neural signatures (low-burden, opportunistic)

While our primary data are behavioral and experiential, we also consider brain and physiological correlates, collected opportunistically and non-invasively when possible. We won't burden patients for this, but in rare cases where someone is interested or when existing clinical monitors can be leveraged, we have some targets in mind:

Where feasible (and only with full opt-in), we could use mobile EEG headbands, pupillometry via a tablet camera, or other passive sensors during awake, calm periods to probe neural signals. No long lab-style tasks—just maybe brief cognitive or choice prompts that can be delivered in a game-like or conversational way, if the patient is willing. Below are some hypothesized signatures, tying back to LoF's proposed control systems:

vmPFC/OFC value boost: In scenarios where a patient is choosing between options that have similar immediate comfort but differ in longer-term "repair" value, LoF suggests the brain's valuation centers (ventromedial prefrontal cortex or orbitofrontal) will assign extra weight to the option that offers closure. A coarse proxy: in EEG, frontal midline theta power may track value integration (without region-level specificity), and in fNIRS or if any imaging is available, vmPFC/OFC hemodynamic activity. We predict an elevated signal for choices that carry repair/meaning potential, and that this elevation scales with $H^{-1}(t)$ (i.e., is stronger when time is shorter). In short, as horizon shrinks, the brain's value coding tilts toward repair options.

ACC/rIFG inhibitory gating: For actions that are now "off the table" (like trying to start a new complex project near death), we expect increased engagement of inhibitory control circuits—dorsal anterior cingulate cortex (ACC) and right inferior frontal gyrus (rIFG). This might manifest in EEG as increased conflict/control-related frontal midline theta power (4–8 Hz) and transient beta activity during inhibitory moments, though such signals are not anatomically specific. Under short horizons, the system should more aggressively prune these options, so these signatures could be stronger.

Insula/autonomic quiet: When a patient opts for a relief-heavy choice (say, choosing comfort like listening to music over something stressful), we expect to see autonomic systems reflect a settling. This could appear as an increase in heart rate variability (e.g. RMSSD) and a decrease in skin conductance, indicating a shift to a calmer physiological state. Essentially, a relief choice should correlate with an "exhale" pattern in the body.

Importantly, we'd check that this isn't solely due to sedation; it should be independent of medication level changes.

DMN "life-review" cohesion: During quiet moments of meaning-making (like if a patient is just lying back thinking peacefully or reminiscing), we might see EEG patterns consistent with the default mode network (DMN) being engaged (noting EEG cannot uniquely localize DMN activity) but in a coordinated way—perhaps increased alpha coherence across frontal-parietal leads and reduced competitive interaction with executive networks. This is speculative, but some literature on end-of-life experiences suggests a shift in brain network dynamics during reflective states. If we get EEG during such periods, we'd look for signs of that integrative "life-review" state (maybe akin to meditative or memory-recall patterns).

All these are optional and will likely only be recorded in a subset of patients who are both interested and physically able, or when clinical monitoring provides analogous signals (e.g. heart rate monitors we can read).

We might also consider naturalistic "tasklets" that are extremely light: for instance, presenting the patient with an imaginary scenario for a minute:

Repair vs. Indulgence choice vignette: "Imagine you have energy for one thing: would you like to call a family member to resolve something, or watch a funny video?" – While they ponder or answer, measure pupillary response or EEG. Hypothesis: under short horizon framing, their decision latency might be shorter for the repair option, and pupil dilation might indicate higher engagement for it.

Call-or-Not prompt: Offer a choice: "Would you like to call person X or do something else (e.g., a trivial activity)?" Track whether the decision to call is taken and any physiological change during the decision. LoF expects that as time feels short, saying yes to the call (the closure act) comes more readily (quicker, with perhaps a calm or resolved physiological response once chosen), whereas choosing the trivial option might come with a lingering sense of conflict (maybe more ACC activity).

Predictions (for neural/physio, high-level):

We expect interactions between compensability (Φ) and horizon in these signals: for instance, the difference in vmPFC signal between a high- Φ option (repair) and a low- Φ option (indulgence) will be larger when the patient perceives time is short. Reaction times or autonomic arousal might diverge similarly (e.g., quick to accept a meaningful act when time is short, slower or more hesitation for trivial acts).

Falsifiers (for neural signatures):

Since these are exploratory, falsification here is more about whether we find *any* consistent signal of QS in the brain. If we deploy these tools and find no patterns—e.g., no detectable difference in brain signals between repair vs. indulgence contexts, ACC/rIFG not doing anything interesting as horizon varies, etc.—then the neural evidence for QS is lacking. If what we observe fits better with a simpler model (say, all signals correlate just with general arousal or fatigue, not with anything about repair vs. neutral), that undercuts the idea that a specialized mechanism is at play.

For example, if we find that vmPFC activity correlates only with immediate reward and not at all with repair potential (contrary to expectation), or that any changes we saw in ACC could be fully explained by general fatigue or medication, that would mean no unique “fairness constraint” signal was found.

Given these would be opportunistic data, lack of evidence here wouldn’t by itself falsify LoF, but positive findings would be a strong bonus. However, if multiple opportunities to find something all show nothing, and especially if a simpler explanation fits, it weakens our confidence that there’s a neurologically distinct QS process in end-of-life decisions.

11.4.5 Nuisance modeling and controls

To ensure we’re not misattributing effects, we pre-plan various covariates and controls:

Medical interventions: We will include measures of total opioid dosage (e.g. morphine equivalents per day) and sedative use as time-varying covariates. If someone is deeply sedated, their behavior and affect might flatten for pharmacological reasons. Our analyses of HCl or act counts will adjust for this. If sedation level is high, we may even exclude those periods from certain analyses (like the centering test, since LoF wouldn’t expect a pattern if the person isn’t really conscious).

Delirium screening: Use a tool like CAM-ICU for regular checks of delirium. If a patient-day is delirious, we know any “acts” might be disorganized or not reflect intentional QS behavior. We’d likely censor or mark delirious periods and handle them separately. They might be excluded from the main analysis and instead treated as interesting failure cases. But for fairness, if a patient was delirious the entire last 3 days, we wouldn’t say LoF failed—rather, that case meets exclusion criteria for the hypothesis (since QS presupposes some coherent agency). We will report how many cases were excluded due to this.

Sleep quality: Using either actigraphy from a wearable (if already being worn, we won’t add one purely for study) or nurse observations of restlessness, we’ll note if someone had extremely poor sleep. This could confound mood and behavior (bad sleep can mimic

or mask LoF effects). We'll include some indicator of recent sleep quality in models predicting affect to partial out general exhaustion effects.

Trait anchors: If available through caregiver report or prior assessment, we might include baseline personality or trait measures that could influence how someone behaves at end-of-life (for example, a resilience score or trait agreeableness). This helps test the “trait-only” rival explanation: maybe people who are naturally more reconciliatory just show these patterns, LoF or not. We will adjust for these or test interactions with horizon to see if LoF effects hold across trait levels.

Staffing/noise log: Because hospital/hospice environments vary, we'll keep track of factors like nurse-to-patient ratio or any day with unusual noise/disruption (construction, etc.). These could impact a patient's comfort and behavior (and our Channel score captures some of it). We may use this in sensitivity analyses to ensure, for example, that a patient's failure to have closure acts wasn't on a day when the unit was chaotic.

Analytic stance: All covariates and model specifications will be decided beforehand (preregistered). We will also run “adversarial” specifications – e.g., a skeptic might say “maybe it's just pain level driving everything.” So we'd run models controlling for pain explicitly to check if horizon effects still show up. We commit to publishing all these alternative model results. Coders tagging events or outcomes will be blinded to things like how long the patient lived or whether we consider their outcome a success or fail for LoF, to avoid bias in qualitative data coding.

The aim of all this is to rule out spurious explanations: if we see what looks like a LoF pattern, we want to be confident it's not because, say, the patient with good family support also just happened to get more pain meds (so was happier), etc. Conversely, if we fail to see a pattern, we want to ensure it's not because noise or missing data hid it.

11.4.6 Sample size and power sketch

Our analyses are somewhat complex (mixed models, within-person effects), but we can estimate needs:

For the behavioral compression/reweighting (Hypothesis A), because these are within-person trends observed over days, the primary information comes from the number of patient-days across patients, though within-person correlation reduces the effective sample size relative to raw day counts. If each patient contributes ~30 days of data (some fewer, some more) and we have 60–100 patients, that's 1800–3000 patient-days. That should give adequate power (>0.8) to detect medium effects (like the entropy drop of 0.25 SD or the act-weight increase of 0.4 we set as SESOI). We'll aim for on the order of 60–

100 patients with sufficient daily data, balancing feasibility (hospice studies are tough to recruit) with power. Simulations will be done to refine this.

For last-window centering (Hypothesis B), this is one data point per patient essentially (their final window outcome vs. their own prior). We'd need enough patients to see a reliable aggregate effect. We estimate needing ~100 patients who have at least two weeks of mood data prior to their last 3 days, to have stable forecasts. 100 gives decent power to detect a moderate paired difference. If effect is smaller, we might need more; conversely, if effect is large, fewer could suffice, but planning for ~100 is reasonable.

For lucid window enrichment (Hypothesis B variant), say we need ~50 patients who had multiple lucid and non-lucid periods (like at least a few of each). Within-person OR detection is strong if each person has repeated observations; 50 people might be enough to see an OR of 2 with >0.8 power in a mixed model (depending on how variable people are). We will get as many as we can but note that not everyone experiences fluctuations; it might be a subset.

Neural/physio data likely will be opportunistic and small-N (maybe a dozen patients?), so those would be treated as exploratory. We won't rely on them for primary confirmations, so they don't dictate sample size except "collect when possible."

We will set interim checks (not hypothesis tests, just feasibility checks) after certain enrollment counts to ensure data quality is sufficient and to update power calcs if variance is different than expected.

We also predefine stopping rules as noted (e.g., if after X patients nothing is trending and CI already excludes SESOI, consider stopping for futility, as per 11.3.7).

SESOIs were already outlined for main effects; those will guide whether we believe an effect is meaningfully non-zero if detected.

11.4.7 Negative controls

To guard against being fooled by artifacts, we incorporate some negative controls – cases where we expect no effect. These help confirm that when we do see something, it's real:

Horizon shuffle test: Within each patient, we can randomly permute the "horizon" labels or timing and re-run analysis to see if spurious effects appear. Any strong pattern should disappear with this permutation. If our analysis pipeline were inadvertently picking up some time trend unrelated to actual horizon, this could reveal it.

Decoy acts: We might include a deliberately neutral act category in the log (something benign like "watched generic TV" or "did crossword puzzle" if applicable) to see if our

coding or analysis falsely flags it as meaningful. The prediction is that truly these neutral acts should not show the same patterns. For example, in reweighting, if everything including decoy acts started showing weighting, we'd suspect rater bias. We expect only Relief/Repair/Meaning acts drive the effects.

Time-of-day effects: We will check if any of our patterns could just be due to day/night cycles or fatigue at end-of-day. For instance, maybe in late evening patients always do less—if deaths often happen at night, last-day data might reflect evening quietness, not LoF. As a control, we might compare morning vs. evening patterns earlier in illness to see if those differences mimic what we ascribe to horizon. If yes, we might need to adjust for time-of-day. Ideally, LoF effects hold even controlling for that.

These controls ensure our statistical pipeline isn't generating false positives and that what we tag as LoF-specific really is tied to the conditions we think (horizon, channels) and not some hidden factor.

11.4.8 Integration rule (what counts as a “hit”?)

To declare that the end-of-life LoF package is supported by the data, we set criteria that span the domains:

Under conditions of high Channel score (good care) and with preregistered analyses and blinding in place, we would require all of the following to consider the LoF end-of-life hypothesis supported:

Behavioral compression and reweighting: The mixed model from Hypothesis A shows $\beta_1 > 0$ for horizon in the entropy model (meaning entropy decreases as horizon shortens) and $\gamma_1 > 0$ in the closure-weight model (meaning closure-weight increases as horizon shortens).

Last-window centering: The group-level analysis for Hypothesis B shows $D_0 - D_f < 0$ (final mean closer to zero than forecast) with at least a medium effect size and statistical significance. In other words, a clear trend that final affect outcomes are nearer neutrality than expected.

At least one neural/physiological signature: We don't need every exploratory measure to work out, but we'd like to see at least one convincing neural or autonomic sign consistent with LoF's QS predictions. For example, maybe we find that the heart-rate variability is significantly higher during high-closure days, or an EEG marker that correlates with horizon on a repair tasklet. If at least one such $\Phi \times H^{-1}$ interaction effect replicates in the small sample we have (or across sites), that gives us added confidence of a mechanistic basis.

If all these conditions are met, we'd say the evidence strongly supports that the Law of Fairness is manifesting at end-of-life, under the conditions tested.

On the other hand:

Failure criteria: If any two of the three domains above (behavioral patterns, affect centering, neural signatures) fail to appear as predicted in replicated analyses, we would *downgrade* the support for LoF in this context. For instance, suppose we do see compression and reweighting in behavior, but no centering in affect and no neural signatures. That's one domain working and two not—this would force a reevaluation, possibly concluding that the law isn't holding strongly, or only part of it is. If all three domains fail (no behavioral pattern, no centering, no neural hint) or if the ultimate ledger measure (see below) diverges, then LoF's end-of-life claim would be essentially rejected for now.

Finally, an ultimate check: we could attempt to estimate the final ledger balance for each person. Using all data, integrate $\hat{F}(t)$ (estimated from HCl or other measures) to get $\hat{L}(t)$ with confidence intervals. If a large fraction of patients have $\hat{L}(t)$ confidently outside a small band around zero (say more than 20% of patients with final balance worse than some $\pm K$, where K is like 0.5 HCl-days, just as an illustration), that's direct evidence against neutrality. We'd incorporate that into the "downgrade" decision as well.

In sum, we've set up the rules so that we're not cherry-picking a single positive finding to claim victory; we want a coherent story across multiple levels. Conversely, if a couple of these tests fail conclusively, we're ready to say the neutral-closure idea is on thin ice, at least as originally formulated.

11.4.9 Minimal toolbox (practical checklist)

To close this technical section, it's useful to summarize what an optimized, *practically deployable* measurement package could look like for hospice settings, based on the above:

One-page Act and Closure Log: A simple paper or tablet form with checkboxes for each of the 10–12 act categories and a space for time-stamps or short notes. Used by staff or family to log what the patient did each day.

4-item HCl card: A small card or app with four 0–10 sliders (or even smiley face scales) for pain, breath ease, mood ease, connectedness. Takes under a minute for patient or proxy to fill, done perhaps morning, afternoon, evening.

0–5 Channel Checklist: A quick daily checklist for the nurse: Was pain controlled? Was a key person present? Quiet environment? etc., to yield the Channel score.

Optional Lucidity 0–3 tick: If patient has fluctuating consciousness, a thrice-daily quick rating of alertness and orientation (or delirium assessment) – only if relevant.

Passive wearables (optional): Only if the patient is already wearing a device for their own interest (like a sleep tracker or heart rate monitor to guide comfort), we might include that data – but we do not introduce wearables solely for the study unless the patient specifically asks.

Analysis plan pre-written: All the code and statistical analysis scripts are prepared in advance and registered. When data comes in, it is processed through this pipeline without fishing.

Null result publication commitment: A document stating that regardless of what we find, the results will be published or shared, especially if outcomes are null or mixed.

Takeaway: By using this minimal toolbox, we can rigorously capture whether variance compression, priority reweighting, and small neural/autonomic correlates are happening as a patient nears end-of-life – all without disrupting care. If these signals persist even after we account for confounds and apply blinding and adversarial tests – and especially if they scale with the horizon and depend on open channels – then we will have captured a public, reproducible fingerprint of a system striving to keep the ledger fair when time is shortest. On the other hand, if careful measurement with this toolbox finds no such patterns, that outcome is equally crucial: it means the LoF end-of-life predictions would need serious revision or abandonment.

11.4.10 Where we go next:

With analysis rules fixed, Section 11.5 addresses interpretation: how to present uncertainty, when to call neutrality, and when to say “inconclusive” with candor.

11.5 Reading Anecdotes: Scripts vs. Signals

Anecdotes are often the first spark for hypotheses—and the quickest way to fool ourselves. End-of-life care is steeped in powerful narratives: *the final rally*, *waiting for permission*, *seeing the light*, and so on. Some of these may be cultural scripts that influence how people behave and how caregivers interpret events. Others might be genuine signals of the underlying balancing process we’re investigating (admissible-set tilt, counterweight bursts, last-window centering). This section provides practical rules for distinguishing story form from functional signal, ensuring that we use anecdotes constructively without being misled by them.

11.5.1 First distinction: function over form

Script (form): A repeated storyline, phrase, or image that people absorb from culture, religion, or media. It’s *how* people think end-of-life is “supposed” to go. For example, “she needed to hold on until her son arrived” or “he saw angels and then he was at peace” are narrative forms that are culturally reinforced.

Signal (function): An observable act or choice that directly affects the hedonic ledger or the probability of closure—regardless of how it’s described. This is *what actually happened* functionally in LoF terms. For example, the patient relaxed and their breathing eased after the son arrived—that’s a Relief function and closure of a relational loop. Or a patient verbalized seeing a comforting figure and then their anxiety dropped—that’s possibly a Relief function (soothing imagery) at work.

Rule: Code what was done or what changed, not the story wrapper around it. If someone says “I saw my grandmother by my bedside,” we don’t log “supernatural vision.” We log the functional outcome: perhaps this experience brought them calm (so we’d code a Relief function, like “soothing imagery” with an outcome of calmer behavior). We explicitly do not treat the anecdote as proof of an afterlife or anything—that’s beyond our empirical remit. Likewise, if a patient says “I’m waiting for my daughter,” and indeed hangs on until the daughter comes and then dies, we code those events: presence of key person followed by death relatively soon after. That is evidence consistent with the “permission” dynamic, but we remain neutral on *why*. The functional code might be: Meaning/closure act achieved, followed by peaceful passing.

In short, we treat anecdotes as data points about actions and outcomes. The meaning the patient or others attribute (e.g. “a soul needs permission”) is respected personally but not baked into our scientific analysis.

(An example: “I saw a bright light and felt no pain after that” → we code that as a moment of subjective relief and perhaps note they appeared less distressed afterward. We don’t code “bright light = going to heaven” or such.)

11.5.2 The three-column note

To keep ourselves honest in interpreting narratives, we can use a simple structured note for any anecdotal report, dividing it into three parts:

Verbatim snippet (≤20 words): Write down exactly what was said or observed, in the patient’s or clinician’s own words if possible. E.g., *Patient*: “My mother is here—I can go with her now.” Or *Nurse note*: “Patient suddenly sang a hymn clearly for two minutes.” No interpretation, just the raw quote or description.

Observable act(s) (with timing): What did the patient *do* or what event occurred around that time? For instance: “Called her sister and said goodbye,” or “After singing, she smiled and closed her eyes peacefully,” or “Morphine dose given at 8:00 pm, family holding her hand.” This focuses on the concrete actions and their time sequence.

Functional tag(s) and context: We then tag the above with our framework: Relief / Repair / Meaning / Neutral, and note any relevant context like Channel score or horizon phase. For the example, we might tag “Meaning (farewell blessing)” or “Repair (apology)” or “Relief (calming vision)” as appropriate. We’d also note if this happened in a lucid window, whether pain was controlled, etc.

By structuring anecdotal entries this way, we convert rich stories into analysis-ready events. We don’t erase the humanity—we keep the verbatim snippet so the richness isn’t lost—but we separate it from what we analyze. This allows us to compile a dataset of events (with tags and context) that can be systematically compared, without being biased by the emotional or scripted veneer each story has.

It’s like having our cake and eating it: we honor the narrative in one column, but the next columns ensure we extract the factual and functional elements for science. This way, anecdotes feed into our analysis without drowning it in subjectivity.

11.5.3 Red flags for script-driven notes

Certain features in documentation or recollection suggest that a narrative might be more script than signal (i.e., the caregiver or patient might be unconsciously following a cultural script in describing events):

Teleology words: If the language implies a purpose or destiny (e.g. “He was *meant* to wait until his anniversary to go” or “It was *her time* because she finished her life’s work”). These phrases indicate we’ve left the realm of observation and entered interpretation.

Clinician projection: Phrases like “He needed forgiveness” or “She couldn’t let go because of guilt” when these are not direct patient quotes. That’s the clinician or family projecting a narrative (the patient never said “I feel guilty,” but the clinician infers it). Such notes might reflect more of the writer’s beliefs than the patient’s actual state.

Absence of acts: The note is full of dramatic language but contains almost no concrete actions. E.g., “A profound peace filled the room as the veil thinned” – poetic, but what actually happened? If we can’t find an act, we can’t code a signal. It’s a red flag that this is pure narrative.

One-sided causality without context: Statements like “After the prayer, she passed peacefully” that don’t mention that, say, morphine was also given 30 minutes prior or that her daughter arrived (other plausible causes for peace). If notes consistently credit a spiritual or narrative cause and ignore practical factors (analgesia, family presence), the narrative may be biasing what is recorded.

Stereotyped rally with no substance: E.g., “Sudden surge of energy” is noted but when you dig, it was just increased wakefulness, with no meaningful interactions, and it might have been due to a medication wearing off. The note calls it a “last rally” but functionally nothing happened except maybe restlessness. That suggests the caregiver may have been primed to see a “surge” because that’s a common story, even though it wasn’t accompanied by closure acts or clear lucidity (so by our definition it wasn’t a true counterweight burst, just a spike in activity).

When we see these red flags dominate a story, we treat that anecdote as cultural context, not data. It can inform us about expectations and biases, but we wouldn’t count it as evidence for or against LoF.

For example, if a nurse’s note says, “Patient had the classic last goodbye dream; therefore, she was at peace,” it’s loaded. We’d extract the facts (“patient reported a dream about saying goodbye”) and outcome (“seemed calm after”), but we ignore the “classic” framing and the “therefore at peace” conclusion without evidence. Red-flag-heavy narratives might be set aside or used carefully with corroboration from data.

11.5.4 Green flags for signal-bearing anecdotes

Conversely, what features suggest an anecdote likely contains a genuine signal aligned with LoF predictions?

Action density: The anecdote is concise and tied to specific behaviors. E.g., “He woke up, clearly asked for his brother, spoke for two minutes, then settled back.” There’s not a lot of flowery commentary—just a lot of *actions* described despite being a short story. High density of doings vs. commentary is a good sign.

Timing clarity: It notes when things happened, especially relative to known events, showing a potential cause-effect. “Within minutes of his pain easing, he started joking with staff” – that’s useful temporal resolution, suggesting relief led to positive social interaction. Or “Daughter arrived at 5 pm; patient died at 5:30 pm after a brief interaction.” This temporal linkage (channels opening then closure act then passing) is exactly what we’d look to test. If anecdotes give timing, we can align them with our timeline data.

Horizon scaling evident: The content of the anecdote aligns with prognosis awareness. For example, “He refused to talk about finances until the doctor said it might be days, then he immediately arranged everything.” That shows behavior changing as horizon info changed – a pretty direct LoF signal that horizon matters. Similarly, “As soon as she heard there was no more treatment, she called each of her children.” These are narrative but strongly hint at horizon-driven change.

Cross-script convergence: If different cultural or personal framings lead to the same functional pattern. For example, one patient prays to be forgiven and then says she’s at peace; another writes letters of apology; another simply tells the nurse she feels resolved about past wrongs. Different form (prayer, letters, conversation), same function (seeking closure on guilt or reconciliation). When anecdotes from varied contexts all point to a similar *function*, that’s green flag that it’s a real signal (the underlying act of making amends, regardless of form).

These “green flag” anecdotes are the ones we want to elevate in our analysis: they become entries in our dataset of events. They often feed into the patterns we specified in 11.4. For instance, clustering of high-yield acts in lucid windows (counterweight bursts) might be first noticed because of an anecdote like “Every time he was clear, he did something important.” A green-flag anecdote like that directly becomes a testable signal.

In essence, green flags = the raw material of good science in narrative form. They point to what data we should quantify and verify.

11.5.5 Expectancy and placebo controls (without cynicism)

We must acknowledge that patients and staff have expectations. Patients might try to be a “good patient” and give a nice ending, or staff might document things in a more positive

light because they want a meaningful story. We don't want to become cynics who dismiss everything as rose-tinted—but we do need controls for expectancy effects:

Blinded coders: We ensure that the people who are coding anecdotal data into our system (the ones deciding if an act was Relief or not, for instance) do not know the context that could bias them. For example, they shouldn't know “this happened on the day the patient died” or “this was supposed to be a last rally.” They just see “Patient sang a song and smiled at 3pm” and code it. This prevents them from confirming narratives unconsciously.

Decoy items in tagging: As mentioned earlier, include some events in the records that should not be meaningful if staff are unbiased. If we see that *every single thing* gets tagged as meaningful by a particular nurse (“Patient asked for water” – meaningful? No, that’s normal – should be Neutral, but an overzealous person might say “Ah, water of life symbol!”). To catch that, we include clearly neutral behaviors in our checklist. If they start showing up as “meaningful” in logs, we know expectancy bias is at play.

Language prompts and training: We train staff in the field to document in a behavior-first way: “Patient did X,” not “Patient must have felt Y or needed Z.” We give gentle feedback like, “This is great, but can you also note what the patient actually said or did around that time?” We encourage notes like “Patient appeared calm after prayer; pain 2/10” instead of “Prayer gave patient peace.” This helps produce more factual documentation. Also, having standard forms like the three-column note naturally nudges people to separate observation from interpretation.

The goal is not to strip away all meaning like cold robots; it's to channel the natural human desire for meaning into *accurate observation* first, then interpretation second. We still allow and record what families and patients *believe* happened (that's the first column verbatim). But our analysis keys off the second and third columns.

11.5.6 The adversarial reading drill

Before we finalize any conclusion that anecdotes support LoF, we do an internal skeptic review. We basically ask: Could there be other reasons for these anecdotal patterns? We run a mental (or literal) checklist:

Fatigue model check: Could the “doing fewer things” simply be because they’re tired, not because they’re focusing? Look at closure weight – did it rise as count fell? If not, maybe it’s just fatigue, not selective LoF behavior.

Availability bias check: Are we only hearing about the “good” stories and not the neutral or negative ones? Maybe staff only report those beautiful moments and not the days of

nothing special. We audit: do we have continuous logs or just cherry-picked highlights? Ensuring consecutive day coverage helps; if it's spotty, ask why (maybe boring days weren't recorded?).

Third-variable check: Did something else explain the nice story? "After prayer, she passed peacefully" – Was it really the prayer or was it that she finally got a bolus of morphine and family presence at the same time? We cross-check Channel data: if all the "peaceful after X" anecdotes coincide with, say, effective symptom management, then X might not be the cause. We incorporate those variables in analysis to see if the pattern holds beyond them.

Rival fit check: Try to fit the anecdotes to a different theory – say, pure trait resilience ("she was always an optimistic person, that's why her end was peaceful") or homeostasis ("the body just shuts down and that looks like peace"). Are there cases the rival explains that LoF wouldn't? Or vice versa? We keep a score: are there any anecdotes where a trait or simple adaptation story makes sense but a horizon/closure story doesn't? Conversely, are there ones that only LoF explains? We consider these systematically.

We basically play devil's advocate with our narrative interpretation, and only what survives this drill without glaring holes goes into our findings. This exercise ensures we don't become evangelists of our theory by over-reading anecdotes.

11.5.7 A compact tagging vocabulary (4 tags, function-mapped)

To streamline and standardize how we label anecdotal (and observational) data, we defined a small set of tags earlier (Relief, Repair, Meaning, Neutral). We can expand those slightly with more concrete sub-tags for ease of use, but still keep it minimal:

Relief – use for acts like: accepted analgesia or comfort measure, listened to music, sought light or touch, embraced quiet/warmth. These all directly relieve distress.

Repair – use for interpersonal mending acts: made a difficult phone call, apologized to someone, forgave someone, de-escalated a conflict, arranged a visit with estranged person. This covers reconciliation or conflict resolution behaviors.

Meaning – use for symbolic or legacy acts: gave a blessing, expressed explicit gratitude, bestowed a keepsake or bequest, did a life review conversation, said goodbye or gave permission for others to move on. Also spiritual acts like prayer or ritual if their purpose is meaning-making rather than immediate relief.

Neutral – use for everything else that doesn't clearly contribute to closure or comfort: e.g., watching random TV, routine medical procedures, small talk about weather, etc.

We instruct that multiple tags can apply to one event if needed (e.g. a conversation might involve both *forgiveness* (Repair) and *gratitude* (Meaning)). But we try to keep each logged “act” focused enough for one primary tag.

We keep tags short and tied to behavior, not theory jargon. Instead of tagging something as “ledger balancing attempt,” we tag it in plain terms like “apology (Repair).”

This compact vocabulary (which indeed maps back to our larger conceptual categories) ensures consistency in coding across observers and sites. It also makes training easier: we can give examples for each of the core tags, and it covers most everything we expect to see.

By limiting tags, we avoid drifting into subjective territory and ensure that when we aggregate data, we’re comparing apples to apples.

11.5.8 How to read common motifs

Let’s apply the above principles to some of the well-known motifs and see how to *functionally* interpret them without the script fluff:

“Waiting for permission”: Script form says a dying person hangs on until given permission by someone to die. Signal test: Did a key person actually arrive or did the patient explicitly get told “It’s okay, you can go now”? If yes, that’s a concrete event. Did the patient’s condition notably change after this (e.g. they died shortly after, or visibly relaxed)? If yes, we tag that as a Meaning function (permission given/received) and Relief (if their distress dropped). If no person arrived or no explicit permission was heard, then “waiting for permission” is just a projection – treat it as script. Essentially: check for the actual presence of the event and its immediate effect. If those are present, this motif yields a signal: relational closure achieved, possibly enabling the final step.

“The last rally”: Script says many have a burst of energy before death. Signal test: Was there actually a lucid period (measured by our lucidity scale or clear behavior) and were high-yield acts done during it? If instead it was just generalized restlessness or temporary improved vitals with no closure acts, then functionally it’s not a QS counterweight, it’s just a surge (and we’d label it Neutral or physiological). So we require evidence of lucidity + closure acts to count it as the meaningful “counterweight burst” that LoF predicts. If it’s just wakefulness without those, we classify it as Neutral (maybe delirium or random fluctuation) even if someone calls it a rally.

“Visions of travel or deceased loved ones”: Script interpretation might be “they are journeying to the next world” or “spirits are visiting.” Signal test: We ignore metaphysics and ask: Did this imagery correspond with a change in the patient’s emotional state or

behavior? If a patient says “I was on a train to a beautiful place” and afterward they seem calmer or they decide “I know what I need to say to my son now,” then we have something: we tag that dream as Relief (soothing imagery) or even Meaning if it inspired an action. If they simply report a vision but it has no observable effect (and they’re not distressed by it), we might tag it as Neutral (interesting but not clearly affecting ledger). The key is whether the content led to any relief or action.

We thus “decode” common motifs by filtering them through functional outcomes. This prevents us from either gullibly accepting every lore or, on the other side, dismissing valuable phenomena just because they come cloaked in mystical language. We find the pragmatic core.

11.5.9 From anecdotes to priors (Bayesian discipline)

All the anecdote-reading above serves another purpose: informing our *priors* for formal analysis. In a Bayesian sense, each credible anecdotal pattern can be the seed of a prior belief about an effect size or probability.

For example, if we compile anecdotes and find, say, 7 out of 10 lucid windows had a reconciliation in them, we might set a prior that the within-person probability (or odds ratio) of that is high. Or if across cultures anecdotes convince us strongly that as H_t shrinks, something consistently happens, we can encode that as a prior distribution on our horizon effect parameters.

In practice, we can treat each tagged act (like each repair attempt or relief choice) as a Bernoulli trial of “did a closure-oriented act happen in this context (given horizon and channels)?”. We can update a Beta distribution as we gather anecdotal evidence even before the formal study data.

Then, as formal data comes in, we have a disciplined way to integrate it: the anecdotes set an expectation (prior), but the data (likelihood) can confirm, shift, or contradict it.

This ensures we don’t throw out anecdotal knowledge, but also don’t let it overpower measured data. If the anecdotes were misleading, the data will push the posterior towards a different conclusion. If they were accurate, the data will strengthen that conclusion and narrow uncertainty.

We will also explicitly test rivals using this framework: for instance, create a simple model (like “chance of a closure act = baseline + something * horizon”) and update with each anecdote as if it were data. If rivals (like “chance of closure act is constant regardless of horizon” vs. “chance increases as horizon shortens”) are on the table, anecdotal

evidence can start shifting the odds between those models even before we run the big study, in a qualitative way.

The point is to treat anecdotal patterns as *informative but not conclusive* – they move our starting point, but the real test is still to come. And if a strong anecdotal prior doesn't pan out with real data, we adjust. This formal approach prevents cherry-picking anecdotes to confirm what we want; it forces us to aggregate them and let them influence our expectations quantitatively.

11.5.10 Composite vignettes that teach without bias

When sharing findings or training staff, we will often use vignettes. To avoid the bias of any one real case (and to protect identity), we should construct composite vignettes – like the ones we gave in 11.2.13. These combine elements from multiple patients into one narrative.

We will make sure to also include negative or neutral examples to avoid a rosy bias: e.g., a vignette of a case where channels were blocked and the patient died in distress, to illustrate what it looks like when LoF *fails* or can't operate. This teaches the boundary conditions.

Each vignette we publish should be structured with the three-column approach: present what was actually said/done (maybe in quotes and description), then what was observed, then an analysis of it. This way, readers can see *why* we interpret it a certain way.

By providing these clear examples, we let others audit our reasoning. It also helps caregivers recognize patterns without the need for theory jargon: they might think, “This scenario looks like that composite vignette – maybe I should ensure the family can visit, because that story emphasized how isolation was the issue.”

The key is transparency in how anecdotes illustrate but do not *define* our conclusions – they support them.

11.5.11 What not to do with anecdotes

Finally, a brief caution list that we adhere to (and recommend to others) regarding anecdotes:

Do not treat emotionally moving stories as proof of anything on their own. They are starting points or illustrative points, not endpoints of evidence. No matter how profound a story of someone finding peace is, it's still one data point.

Do not use anecdotes to push any ideological or spiritual agenda. E.g., we won't say "Because of these stories, you should believe in the afterlife or the power of prayer." Our role is to observe and support patients, not to proselytize any worldview, even a secular one.

Do not correct or argue with patients' or families' interpretations. If a family says "It was definitely our prayers that helped him let go," we don't have to agree, but we absolutely do not say "Actually, it was the morphine." We code internally what we need to (prayer event, morphine given), but we respect their narrative as *theirs*. We don't impose our interpretation on those living the experience.

Do not over-fit your theory to every anecdote. One family might have a very unique ritual or a patient might have a very idiosyncratic way of coping – we must be careful not to create a special sub-clause in our theory just to rationalize it. If something truly doesn't fit, we acknowledge that as a limitation or an open question, rather than bending the theory beyond recognition. One-off anecdotes should not unduly warp a general model – unless of course they accumulate into a pattern.

In other words, we keep a balance: treat anecdotes with humanity and curiosity, but always corroborate with data and broader patterns.

11.5.12 A tiny field checklist (fits on a badge)

To help busy hospice staff assist in data collection without having to remember all this, we can provide a quick-reference checklist – something that could even fit on the back of an ID badge:

Write the act. (What did the patient *do* or say? "Patient hugged sister and said 'forgive me.'")

Time-stamp it. ("After morphine dose at 2 pm" or "9:30 am" etc.)

Tag the function. (Relief/Repair/Meaning/Neutral – pick the best fit or multiple if needed.)

Note channel context. (Pain level, who was there, environment – e.g. "pain 2/10, daughter present.")

Skip the sermon. (Avoid adding "therefore this means..." in the note. Just observe.)

These five steps ensure that what is recorded can be used scientifically *and* is respectful. It basically encapsulates everything: factual record, timing, interpretation kept separate.

By turning anecdote gathering into something closer to data gathering—without losing compassion—we ensure our study remains rigorous and that any conclusions about LoF will stand on solid ground, not just beautiful stories.

11.5.13 Where we go next:

No claim survives without fail patterns. Section 11.6 specifies what findings would count against end-of-life neutrality and how to publish those results quickly and respectfully.

11.6 Fail Patterns at Terminal Closure

A scientific law earns its credibility not just by predicting what will happen, but by specifying what would count against it. In this section, we list the observable, preregistered failure patterns that, if seen (especially under good care conditions and clear measurement), would force us to significantly doubt or reject the Law of Fairness's claims about end-of-life. Each Fail pattern includes how we'd recognize it (operational criteria), what else could explain it (differential diagnoses to rule out), and how we'd report it if it occurred.

11.6.1 Fail pattern: no variance compression, no priority reweighting

Prediction recap: If LoF holds, as the horizon H_t shrinks (and if channels are open), we expect to see the patient's behavioral repertoire narrow (fewer distinct acts, lower act entropy) and their choices tilt towards relief/reconnection (higher proportion of closure-focused acts).

Fail pattern: Across a broad sample of patients (with adequate symptom control and social support – say, Channel score ≥ 4 out of 5) we observe *neither* of these trends:

No entropy decline: Act entropy does not decrease from long-horizon periods to short-horizon periods. It might even increase or just remain flat. In practical terms, patients continue to do as many types of things (or more) when death is imminent as they did earlier.

No closure tilt: The fraction of acts that are Repair/Relief/Meaning does not increase (or even decreases) as the end nears, once we adjust for other factors. In other words, patients in their final days engage in no more closure-seeking behaviors (proportionally) than they did weeks prior; they might even engage in fewer.

Decision rule: Using the mixed models from 11.4.1, this would manifest as the horizon coefficient $\beta_1 \leq 0$ for entropy (no compression) and $\gamma_1 \leq 0$ for closure-weight (no reweighting or negative reweighting), with confidence intervals excluding any positive effect of meaningful size. We'd require this outcome in two independent cohorts (or a large multi-site cohort split in half for replication) with sufficient precision (standard errors small relative to SESOI) to declare this a real failure. In essence, if two well-conducted studies show no sign of narrowing or tilting when there should be, LoF's behavioral predictions fail.

Differential diagnoses to rule out: Before concluding LoF is falsified, we'd check if perhaps these patients had hidden barriers (maybe “Channel score ≥ 4 ” wasn't capturing some qualitative barrier). We'd also check if documentation was incomplete (maybe

they did fewer closure acts but staff didn't log them diligently?). We'd ensure that high-channel days truly dominated the sample. Also consider if heavy delirium or sedation muddled the pattern – ideally we already excluded those cases. If after considering these, the result stands (i.e. even in clear-minded, well-supported cases no compression/tilt), then it's a genuine failure of the LoF prediction, not just a context issue.

This Fail pattern, if confirmed, would strike at the core behavioral claim of LoF. It means people do not necessarily focus or prioritize emotionally balancing acts even at life's extreme, which would seriously challenge the universality of the law.

(Fail pattern box 11.6.1: no narrowing or tilting – e.g., patient continues planning new projects till the end, and doesn't particularly seek comfort or closure more than before.)

11.6.2 Fail pattern: last-window divergence rather than centering

Prediction recap: LoF asserts that, with channels open, the final conscious days' affect will gravitate toward neutral (no large unresolved surplus of pain or pleasure). Our expectation was $D_0 < D_f$ (actual final mean closer to zero than forecast).

Fail pattern: Using our forecast vs. outcome method (see 11.4.2), we find that the final window's affect not only fails to center, but in a significant number of cases it stays as far or farther from neutral as predicted by prior trends. Concretely:

The mean distance from zero (D_0) is not smaller than the distance from forecast (D_f) for most patients. In many cases, it's about the same or worse.

At the group level, $D_0 - D_f \geq 0$ (no improvement towards neutral) and the confidence interval around the mean difference is on the positive side or contains zero but excludes any meaningful negative difference.

In other words, patients end their lives with an affect balance that is no closer to neutral than what you'd expect if whatever suffering or elation they had just kept going.

We might also detect outright divergence: some cases where final affect is even more skewed than the trend predicted (e.g., someone was on a downward trend and actually ends even lower than extrapolated).

Decision rule: If two independent samples (meeting our power criteria) both show D_0 is not significantly less than D_f – say, a paired analysis gives a difference of roughly 0 and a tight CI, or even >0 – then centering is not observed. Particularly damning is if the combined data show $D_0 - D_f \geq 0$ with narrow confidence intervals at the group level. We would exclude cases that *cannot* center due to immediate physiological catastrophe

(sudden massive stroke, etc., which we design out of the analysis). So this is under conditions where centering *should* have a chance. If even then it doesn't appear, that's a fail.

Differentials: We'd ensure these patients had "adequate measurement density" – meaning we actually had enough data to make a solid forecast (≥ 2 weeks of data prior as planned). If we only had sparse data, maybe our forecast was off. So we'd repeat analysis requiring high data completeness. If centering still fails, it's real. Also rule out cases where death was so sudden that maybe no psychological "last window" occurred – we're focusing on conscious last windows. If after filtering to those, still no centering, LoF's closure principle fails here.

Other confounds: maybe our model wasn't good – but we gave benefit of doubt by comparing to personal trend. If personal trend was steep negative and they continued negative, that's clearly a fail of counterbalancing.

Thus, this Fail pattern indicates that, even with time and comfort, the person's trajectory didn't bend toward okay – it just kept going as it was, or worse. That would seriously undercut LoF.

(Fail pattern Box 11.6.2: No "equilibration" – e.g., patient who was very depressed remains very depressed to the end, no sign of emotional rebound, despite good care.)

11.6.3 Fail pattern: no enrichment of high-yield acts in lucid windows

Prediction recap: When a patient has fluctuating lucidity, LoF predicts that during the clear periods, they should preferentially engage in closure/repair acts (our "counterweight bursts").

Fail pattern: Within-person, there is no difference in the rate of meaningful acts between lucid and non-lucid times. Formally, the odds ratio of a high-yield act in lucid vs. non-lucid intervals is about 1.0, with tight confidence intervals (or even <1.0).

Translated: a patient is just as (un)likely to do something significant when confused as when lucid. There's no special "use of clarity" for closure.

Decision rule: Pool data from all patients with fluctuating courses. If the combined OR for, say, reconciliation attempts or expressions of meaning in lucid windows is ~ 1 and its CI excludes any OR >1.5 (our threshold), that fails the prediction. We'd adjust for family presence etc., as planned. If after that the effect is null, it counts. Specifically, if we see $OR \leq 1.1$ consistently after matching for presence of visitors and pain (with, say, CI like [0.9, 1.3]), we conclude no enrichment. We'd likely require this across two independent datasets or a large one, as usual.

Differentials: Check misclassification: Maybe what we called “lucid” wasn’t actually lucid (e.g., a patient might be quietly delirious but seem lucid). We rely on delirium assessments to ensure clarity periods were truly clear. Check confounds: perhaps in some cases, every time they were lucid, no family was around, so they couldn’t do much – but we would have adjusted for that in analysis. Also ensure that sedation or exhaustion didn’t immediately follow lucid windows cutting them short (making it unfair to expect an act). If none of these issues explain it, then indeed no pattern is there.

This Fail pattern means even when patients get a moment of clarity, they often don’t particularly do anything special. That would imply the “counterweight” idea is not a general rule – maybe just anecdotal in some, but not systematic.

(Fail pattern box 11.6.3: in patient data, lucid times had same behavior as other times – e.g., patient mostly just rests or chats idly even when alert, with no surge of closure activity.)

11.6.4 Fail pattern: neural and autonomic nulls after nuisance modeling

Prediction recap: We posited subtle neural correlates (like vmPFC signals, ACC gating, etc.) and physiological signs (HRV increase, etc.) associated with closure-oriented decisions, especially scaling with short horizons.

Fail pattern: After accounting for obvious factors (analgesia levels, delirium, general arousal), *none* of the expected neural/physio signatures show up. Specifically, any initial interesting finding vanishes when we control for nuisance variables, or even flips sign or is indistinguishable from random fluctuations.

For example:

The $\Phi \times H^{-1}$ interaction effect (e.g., the difference in neural signal for meaningful vs. neutral choices under short horizon) is effectively zero in all regions measured.

Patterns we thought might indicate QS (like ACC activity for inhibited acts) either don’t differ from baseline or are equally present in control conditions (e.g., they also occur in tasks with no horizon difference, meaning it was just a generic effort signal).

Decoy contrasts (like neutral vs. neutral choices) produce similar “effects” as our supposed meaningful contrasts, implying our measures had no specificity.

And importantly, any slight signals could be better explained by alternative models (like fatigue, or simple reward/pain processing) once we factor those in.

Decision rule: We likely won’t have huge sample sizes here, but if in, say, two opportunistic datasets or one reasonably sized one, every targeted neural/physio

indicator yields null results (and we make sure we had enough data quality to see something if it were there), that counts. For instance, if we have 20 patients with EEG during choices and we find no consistent change in frontal theta or any other marker when comparing high-Φ vs. low-Φ under short horizon, that's a fail for that domain. If similarly HRV shows no pattern at act choices after controlling for sedation, etc., that's another fail. We bundle this: if all measured QS signatures are null or inconsistent, we consider LoF's mechanistic support lacking.

We also apply correction for multiple comparisons – if one out of 10 measures showed something but could be a fluke, we wouldn't count that as success unless it replicated. So basically, evidence of absence across the board is the criterion.

Differentials: Perhaps our measurement was too sparse or low-quality (e.g., many EEG artifacts, few patients consenting). We'd acknowledge if power was limited. This Fail pattern might be considered weaker evidence against LoF (behavior and affect weigh more), but it still matters for mechanism plausibility. Also consider that maybe neural signatures require specific tasks to elicit; our opportunistic approach might have been too gentle to detect anything. We factor that in. But if we gave ourselves a fair chance and saw nothing, we log it as a fail in that area.

A complete absence of physiological correlates, especially if behavioral effects also faltered, would suggest LoF doesn't have the distinct "control signals" we expected – maybe it's just a conceptual pattern, not a distinct brain process.

(Fail pattern box 11.6.4: Eg., EEG during end-of-life choices looks no different than random noise or than EEG during non-meaningful tasks, once medication effects are accounted for.)

11.6.5 Fail pattern: menu expansion toward novelty near closure

Prediction recap: LoF says admissible menu shrinks and focuses on closure/comfort as the end nears. The opposite would be a patient expanding into new, non-closure pursuits (which LoF would deem risky/uncompensable moves) even in their final days.

Fail pattern: As time grows short (in patients with good symptom management), we observe *more* distinct novel pursuits or a resurgence of interest in new, non-relief activities – essentially the menu broadens or at least doesn't narrow:

Patients take up new hobbies or projects that have no clear closure or meaning function (and not legacy projects – truly novel indulgent things).

The rate of reconciliation attempts does not rise – perhaps even falls compared to earlier in illness (contrary to expectation of focusing on relationships, they might withdraw or stick to routine).

Act entropy stays steady or even increases (they dabble in a wider variety of activities) *without* a corresponding rise in closure weight.

Decision rule: We gauge this qualitatively and quantitatively. If we saw, for instance, act entropy trending up in short horizon and the proportion of closure acts flat or down, that's expansion. Or if multiple patients started brand-new ventures (unrelated to closure) in their last days and that pattern is systematic. The mixed model would show $\beta_1 < 0$ for horizon effect on entropy (meaning horizon shortens, entropy doesn't drop or goes up) and $\gamma_1 \leq 0$ on closure weighting (no tilt). That overlaps with Fail pattern 11.6.1, but here specifically emphasizing the presence of *novel, non-closure acts*. Perhaps easiest to capture by counting “novel acts” – things the patient never did before – appearing in the last week. If that count is higher than earlier weeks systematically, that's a fail.

We might require evidence across at least two sites or cultural contexts to avoid a local custom misinterpreted (e.g., some cultures have a ritual of starting a new thing as a form of closure – but if it's functional closure, we'd tag it differently).

Differentials: Could some “novel” acts actually serve meaning unbeknownst to us? (For example, writing a poem might look new but could be legacy.) We'd double-check function – only count clearly non-closure novelty like “planning a vacation” with no chance of happening. Also, ensure these aren't delirious or random due to meds. If cognitive impairment is absent and they rationally start something new, that's significant. Also rule out that maybe they always had wide interests (trait) – but an increase is an increase.

If confirmed, this pattern means the system isn't inherently limiting scope at end – some folks might still branch out spontaneously, contradicting a strict LoF.

(Fail pattern box 11.6.5: Eg., a patient in final days suddenly begins learning a new language or starting a business plan with no clear link to closure – horizon short but menu broadens unexpectedly.)

11.6.6 Fail pattern: asymmetric worsening with channel opening

Prediction recap: Opening channels (pain relief, family presence, etc.) should enable closure and *improve* net affect (or at least not harm it). LoF doesn't anticipate that making things better (in care terms) would make the ledger worse – that would be bizarre under the “system strives for balance” idea.

Fail pattern: After a clear improvement in conditions (e.g., effective analgesia administered, estranged family member arrives), we observe an *unexpected deterioration* in patient's mood or closure behaviors:

Despite pain being controlled or a loved one arriving, the patient's HCl scores worsen and they initiate fewer closure acts than before (going against our Relief priority expectation).

Essentially, the very act of opening a channel correlates with a drop in closure activity or an increase in distress, contrary to what LoF would predict (which would be that opening channels either has positive or neutral effects, not negative).

In statistical terms, if we do a within-person analysis of "before vs. after channel improvement" and find that on average mood got worse or closure act rate fell in the days after improvement, consistently across patients, that's a reversal of expectation.

Decision rule: This could emerge from our models if a Horizon × Channel interaction turned out negative (meaning when channels improve under short horizon, outcomes got worse). Or simply observationally: if in a significant minority of cases we see e.g. "once pain was relieved, patient became more agitated or depressed and withdrew." If repeated enough to be significant, that's a Fail pattern.

We'd likely look for at least a handful of well-documented examples. If we can't attribute those to something else, and especially if any aggregated metric shows a paradoxical effect of channel improvement, we'd count it.

Differentials: Perhaps removing pain allowed them to confront reality and that caused psychological distress – i.e., analgesia removed distraction, revealing unresolved issues that then *didn't* get resolved but just upset them. That's actually a dynamic LoF might acknowledge (if too late to resolve issues, relief might expose pain without time to fix). But if that's common, it undercuts "relief helps." We'd ensure no confounding medical issue (maybe treating pain had side effects that caused delirium or something). If not, and it's genuinely an emotional/spiritual downturn triggered by better conditions, that's important.

This pattern would suggest that the QS "tilt" may not engage properly even if conditions are right – possibly pointing to psychological complexities outside LoF's simple model.

(Fail pattern box 11.6.6: Eg., once a patient's severe pain was alleviated and his family gathered (seemingly ideal conditions), he became noticeably more despondent and refused further interaction, contrary to expectation.)

11.6.7 Fail pattern: positive anecdotes without behavioral correlates

Prediction recap: LoF's evidence should come from acts and measurable outcomes, not just stories. If everyone *talks* about how peaceful someone was, but nothing in the data shows it, that's a bad sign (it suggests a narrative placebo effect rather than real signal).

Fail pattern: Sites or cases where there are abundant “good” narratives but our logs show no corresponding behavioral or quantitative evidence:

Many patients are described in qualitative terms as having found peace or meaning, but our act logs show no reconciliation attempts, no notable affect change, no particular closure acts. For instance, charts full of phrases like “died peacefully” yet HCI readings remained high distress until the end, or no concrete closure events recorded.

Essentially, the staff or family narratives are glowing, but the numerical data (act counts, mood scores) are flat. This discrepancy suggests that expectancy or cultural scripting might be painting a false picture of balance.

Decision rule: We would identify this by looking at units or studies with high “narrative density” (lots of anecdotal emphasis on positive experiences) but where our measured indices (like proportion of closure acts, final-week HCI) do not budge or differ from other units. If, say, one hospice is very story-driven and always reports nice anecdotes but when we crunch the numbers those patients didn’t actually have better mood or more closure acts, that indicates a systematic bias in reporting.

We likely treat it as a methodological failure rather than a theoretical one initially – i.e., this triggers method correction (retraining staff, etc.). But if even after trying to correct, the pattern persists (all talk, no data), then either LoF is not manifest or not measurable in that context.

Remedy vs. failure: The first step on seeing this is to suspect measurement error (maybe staff are not logging acts diligently because the narrative “feels” enough). We’d implement a fix: perhaps a checklist to ensure acts are logged, or double-blind assessments. If after that the discrepancy persists (still glowing narratives but data says otherwise), then it’s likely that those narratives were inflated. If despite remedial training the issue remains, it means either LoF isn’t real and people are just overlaying nice stories, or our tools can’t capture it. Either way, it diminishes support for LoF. But ethically and scientifically, we treat it as a sign to improve methodology first.

Only if it “persists after retraining and blinding” do we count it as evidence *against the theory*, in the sense that maybe what we thought were LoF signals are actually artifacts of hopeful storytelling.

(Fail pattern box 11.6.7: Eg., every case is described by staff as meaningful and balanced, yet objective logs show patients died with high pain and few interactions – indicating a feel-good bias.)

11.6.8 Fail pattern: trait-only explanations beat horizon models

Prediction recap: LoF effects should scale with horizon (a situational factor) and not be fully explainable by stable traits (like some people are just resilient, some not).

Fail pattern: When we formally compare models, the ones that use only time-invariant individual differences (personality, spiritual faith, etc.) or trivial processes (like eventual fatigue) explain the data as well as or better than our horizon/QS model:

In model comparison, a model that has no horizon term but includes, say, a “resilience” score, or just assumes everyone tends to baseline, could fit the mood trajectories or behavior frequencies just as well as LoF’s time-sensitive model. If adding horizon parameters doesn’t improve predictive accuracy (or worse, makes it worse), that’s a fail.

Statistically, this could be seen if the Bayes factor > 10 or $\Delta\text{AIC}/\text{BIC}$ clearly favor a simpler model where only traits or a generic adaptation effect (not dependent on time to death) drive outcomes. For example, maybe all that happens is people with high trait optimism die happier, those with low trait optimism die sad – no special end-of-life dynamic. If our analysis finds that after accounting for such traits, horizon length adds essentially no predictive power, LoF is in trouble.

Decision rule: We would specify a priori some rival models (as per 11.3.10 adversarial analysis). If across our data the LoF-based model is not distinguishable from (or is worse than) a trait-based model by conventional criteria (e.g., WAIC/LOO cross-validated log-loss; if WAIC/LOO and log-loss don’t improve with LoF terms compared to a trait-only model), then LoF is not necessary to explain the patterns.

Alternatively, if the horizon terms in our mixed models consistently drop to ~ 0 and the individual random effects account for variance instead, that also hints trait/individual differences dominate.

Differentials: It could be that our sample had restricted range of horizons or traits. But if sufficiently varied and still trait wins, that suggests what we’re seeing is more “who the person is” than “the situation of nearing death.” Also consider combined models: if adding horizon doesn’t significantly improve a trait-based model (and the threshold for “no improvement” is pre-specified, like Bayes factor $< 1/10$ in favor of horizon model or $\Delta\text{AIC} < -10$ favoring simpler model), we judge that horizon is not pulling weight.

If this Fail pattern is observed with measurement invariance established and channels logged as open, it would falsify LoF. A universal claim does not permit “context boxes” once controls are satisfied.

(Fail pattern box 11.6.8: Eg., we find that the only consistent predictor of end-of-life mood and actions is the person’s pre-existing coping style; knowing it’s “end of life” per se adds no predictive value once you know the person.)

11.6.9 Fail pattern: cross-cultural non-replication at the functional level

Prediction recap: LoF claims to be a fundamental dynamic that should appear across human cultures, even if expressed through different forms. Functions (relief, repair, meaning) should converge across cultural settings when measured appropriately.

Fail pattern: In at least two culturally distinct settings (say, hospice in a Western country vs. an Eastern country, or secular vs. religious contexts) with robust community oversight and carefully invariant measures, we do *not* see the LoF patterns:

No increase in Relief/Repair/Meaning act frequency as death nears in one or both cultures, even though in another it might appear. Or patterns actually reverse: e.g., in one culture, people get more withdrawn (fewer closure acts) as death approaches despite good care, whereas another culture shows the opposite.

Alternatively, we try to establish measurement invariance (configural, metric, ideally scalar) for our scales across cultures (pain, mood, act categories). If we manage to do so (meaning we’re truly comparing apples to apples), and then find that our key outcomes (like last-window centering or act tilt) happen in culture A but not in B (or significantly weaker/absent in B), that’s a failure of universality.

Decision rule: Achieving scalar invariance is tough; even metric invariance might suffice for comparison. But let’s say we reach a point where the instruments are reasonably comparable (calibrated anchors as in Chapter 8). If under those conditions one site shows no LoF effect where another does, especially if sample sizes are adequate, that counts as a fail. We’d likely need replication of that (e.g., two non-Western sites show a deviation) to be sure it’s not just one odd sample.

Quantitatively: if a model including culture interaction shows a significant interaction (patterns present in one culture significantly different than in another, not explained by other factors), and especially if in one culture the effects are null or opposite, we consider LoF not general. If scalar invariance holds and yet, say, culture X’s patients do not show final neutralization at all (mean final HCI maybe stays very negative) while culture Y does, that’s what we’re capturing.

Differentials: Ensure that any lack of pattern isn't due to our failure to fully adapt measures. Maybe we missed a subtle cultural expression – but we tried to involve local advisors etc. If after all that, nothing emerges, could be genuine cultural moderation of LoF. Possibly human behavior is more culturally shaped in this respect than LoF allows.

If indeed the functions don't converge (like maybe one culture's way of dealing with death doesn't aim for emotional neutrality at all), then LoF might not be a human-universal law, just a common pattern in some contexts.

(Fail pattern box 11.6.9: Eg., in a study, U.S. patients show the expected balancing behaviors, but Japanese patients (with appropriate cultural measures) do not – their emotional trajectory and acts follow different norms even under good care, contradicting universality.)

11.6.10 Fail pattern: ledger divergence despite prolonged, clear end windows

Prediction recap: LoF's ultimate test is that if a person has enough time and support at end-of-life, their cumulative ledger $L(T)$ should end up near zero (within tolerance K). If someone has weeks of clear-minded life with good care right up to death, there's basically no excuse—LoF would predict neutral closure should happen.

Fail pattern: Identify individuals (and more than a trivial number of them) who had what we'd consider ideal conditions for closure – e.g., at least 2–3 weeks of conscious, supported dying process (hospice, pain managed, family around), and yet their net affect kept drifting away from zero (either positive or negative) and never came back:

E.g., a patient with extended time remains heavily negative in mood and doesn't rebound, or perhaps even accumulates more suffering, with no balancing positive period. Or someone extremely positive never experiences any grounding humility or downturn.

Essentially, their $L(T)$ trajectory over those final ≥ 10 days shows a monotonic drift away from zero (either direction) without any counterbalancing. And this is not due to uncontrolled symptoms (we set "clear end window" meaning no new major medical crises, just the natural decline).

Decision rule: This is a more individual-case pattern. If we find, say, ≥ 30 such cases across sites (or a significant proportion of carefully studied cases) where despite ample time (≥ 10 observation days with high channel scores) the trend of their daily HCl or cumulative ledger is significantly away from zero (and maybe ends $>K$ away, where K is our neutrality bound like $\pm 0.5 \text{ HCU}\cdot\text{days}$ or similar), that's a direct strike against LoF. Two independent samples each with multiple such cases would seal it.

We'd illustrate it by state-space residuals: if the residual (actual minus predicted neutral line) trends downward (or upward) consistently in final days for a number of patients.

Differentials: Exclude cases where, say, they had refractory symptoms we couldn't alleviate (that violates "clear window" assumption). Only count those where, medically and socially, things were as good as they can be. If still ledger goes off track, LoF fails for those cases. If it's, say, systematically in certain conditions (maybe a certain disease like neurological illnesses where it doesn't happen), that's important too.

This pattern, if common, is basically the nightmare for LoF – evidence that even with everything lined up, some people's experiences end unbalanced.

(Fail pattern box 11.6.10: Eg., a patient with three well-supported weeks to live sinks deeper into despair each day with no upturn, dying with a heavily negative ledger despite everyone's best efforts – a clear neutral-closure failure.)

11.6.11 Fail pattern: documentation asymmetries mimicking failure (guarded)

Predicted artifact: This is a nuance – we anticipate that sometimes, staff may document negative events (pain crises, agitation) more diligently than quiet, "nothing happened" moments of closure (which might not get charted because they aren't clinical problems). This could artificially create the appearance that patients had more negative than positive moments (because positive ones went unrecorded).

Artifact Fail pattern: If our data shows what looks like a failure – e.g., no closure acts – but on auditing we discover that staff simply weren't writing down certain kinds of events (like they always note when pain is bad, but rarely note when patient was calm listening to music for an hour), then the apparent lack of closure could be an illusion of documentation.

We call this out because if, say, one site's nurses are too busy to fill in our comfort logs except when something goes wrong, the dataset will be biased towards failure.

Handling: This triggers a measurement remediation, not an immediate theory rejection. We would implement spot-checks or completeness audits. For example, ensure that for each patient-day we have some entry, and if not, that itself is flagged as a data issue.

If after improving documentation practices, the pattern of failure remains, then it's real. If it disappears, then it wasn't a true fail of LoF, it was a fail of data capture.

So while this "Fail pattern" doesn't challenge LoF per se, we include it to remind us to be careful: an imbalance in recorded negatives vs. positives could mislead.

This is why having families or volunteers help note positive moments might help balance the record.

11.6.12 Aggregated decision framework

After enumerating these scenarios, how do we weigh them overall? We propose a simple classification of outcomes for the end-of-life test of LoF:

Supported: If we observe the key positive signatures (compression and reweighting *and* centering *and* at least one neural/physio sign) consistently under high-channel conditions and none of the serious Fail patterns manifest (or only rare exceptions), we declare the end-of-life predictions supported. Minor anomalies might be noted, but the law stands strong.

Under strain: If exactly one of the three major domains fails repeatedly while others hold. For example, maybe behavior patterns appear but centering does not (everyone dies with a slight skew). Or centering happens but neural signals never show (less severe). We'd then say the LoF end-of-life hypothesis is *under strain* or partial – maybe needs revision in that aspect, but not entirely broken. We'd concentrate on fixing or explaining that domain.

Downgraded to tendency: If any two domains fail reproducibly (say, no behavioral pattern *and* no centering, even if maybe a neural hint or two exists; or behavioral and neural fail but centering somewhat holds), then we'd downgrade LoF from a “law” claim to perhaps just a tendency or pattern that isn’t reliable. Essentially, more than one pillar crumbled, so the strong form of the hypothesis doesn’t hold up. We might then integrate with rival theories or restrict scope.

Rejected (EoL scope): If all three primary domains fail in our tests, or if the ultimate ledger test fails (like many people with prolonged windows still diverge), then the Law of Fairness does not hold in the end-of-life context. We'd have to either abandon or profoundly rethink the theory in light of that. It might still hold in other contexts (maybe in dreams or daily life balances) but not at EoL, which falsifies a core prediction it made.

All such determinations must be based on pre-registered thresholds, adversarial analyses, and full publication of null results – we won’t cherry-pick or retroactively move goalposts.

11.6.13 Reporting protocol (how to publish a fail)

If one or more Fail patterns appear, we have an ethical obligation to report it transparently (and we've committed to that). Here's how we will handle dissemination if LoF predictions fail:

Lead with care quality: In any report or paper, first detail the quality of care context (Channel scores, symptom control levels, how dignified the setting was) so readers know these results aren't due to obvious neglect. E.g., "In X patients with excellent palliative care (pain mostly ≤3/10, family present 90% of days)... we observed [Fail pattern]." This sets the stage that we truly tested LoF under decent conditions. If care was suboptimal, we'll say so and frame the conclusions carefully (like, LoF may fail if care fails – which is a different message).

Show the models and data: Provide the statistical models, plots of individual trajectories, and even de-identified raw data (within privacy limits) so that others can verify the fail. If LoF is failing, we need to convince the community with evidence. Full code, anonymized data points, etc., will be shared.

Offer rival fits: As part of the report, we will include how simpler or alternative models fared. For instance, "A trait-only model actually explained these results better, here are those metrics" or "The data were consistent with random variation or with a constant risk model." This shows we tried other explanations and they perhaps fit.

Call scope limits: We'll explicitly delineate what part of LoF is failing. E.g., "These results do not necessarily invalidate LoF in other domains (like dreams or shorter-term compensations), but they do show that the end-of-life neutrality claim is not supported under conditions XYZ." We clarify if the entire LoF concept is in question or just this aspect. Maybe the law holds generally but not in final extreme situations – we'd say that if so.

The bottom line of this section: strong theories state what would prove them wrong, and we have done so above. If those red lines are crossed, we will report that openly, revise or abandon the claim, and share all relevant data. No hiding behind excuses.

Why making failure easy strengthens the claim: By having clear fail criteria and adhering to them, we ensure that if LoF survives these tests, it's truly robust. And if it doesn't, science advances by knowing what's not true. In either case, patients benefit: we either validate a principle that could improve care or we discard a mistaken notion that might misdirect efforts.

If our Law of Fairness emerges intact through all these attempted falsifications – in multiple cultures, under blinded scrutiny, with adversarial collaboration – then those remaining signals demand attention. They will be leaner (stripped of sentiment) and cleaner (replicated under stress tests), and hence harder to ignore as real evidence of a deep regulatory principle.

Having tested the law in the most humane yet uncontrolled setting (hospice), we now move to controlled experiments. In the next chapter, we step into the lab, where we can manipulate time horizons ethically and repeatedly to probe the mechanisms behind these patterns. We will design tasks that mimic end-of-life decisions in miniature, and even perturb the brain's control centers noninvasively, to see if we can induce or disrupt the balancing behaviors on demand. This will further challenge the Law of Fairness under rigorous conditions and help us pin down causality. The journey continues from the bedside to the bench, always with the question: if there is a fairness constraint in conscious experience, how can we know for sure?

11.6.14 Where we go next:

End-of-life tests conclude Part VI's humane evidence arc. Next comes Chapter 12, where we move into controlled horizon tasks and causal probes, making the mechanism carry its weight under time-pressure without ever compromising comfort.

Chapter 12 — The Lab Bench: Horizon Tasks and TMS

Hospice is where horizon effects occur naturally; the lab is where they become *manipulable*. This chapter turns the end-of-life predictions into controlled experiments that any well-equipped behavioral neuroscience lab can run – without invoking metaphysics and without putting anyone in actual peril. We do three things in this translational move: (1) Induce short-term horizons in a safe, reversible way, (2) Measure how people’s choice dynamics respond (especially concerning compensatory vs. indulgent options), and (3) Perturb the brain’s suspected control hubs to test for causality.

Why is this important? Because to truly convince ourselves (and the world) that a fairness mechanism exists, we should be able to *trigger* its signature on demand in a lab setting. We can’t wait around for life-and-death situations for every test – instead, we simulate the pressure of a closing horizon in miniature. Think of it as creating “mini endgames”: scenarios where time is running out, stakes are defined, and we can see if the decision policy tilts in the compensatory direction LoF predicts. Crucially, in the lab we can also apply causal probes – like nudging specific brain regions with transcranial magnetic stimulation – to see if disrupting the supposed fairness controller (the Queue System) alters behavior in the predicted way. If it does, that would support a causal role for those regions in the observed horizon-sensitive behavior, subject to the usual caveats about network-level and off-target effects.

Here’s the plan in a nutshell:

- Induce horizons (minutes to hours): We create experimental conditions that make participants *feel* like they have either a short time or ample time to achieve some goal. This can be as simple as adding a countdown timer (“only 5 trials left!”) or framing a task as “last chance” versus “you can try again later.” For example, one group might be told they have one minute to earn as many points as possible (short horizon) while another has no time limit (long horizon). We can also simulate a “life horizon” in a game narrative (e.g., a fixed number of rounds before retirement) versus an open-ended game, ensuring that any framing differences are matched on stakes and reward structure.
- Measure admissible-set dynamics: We design tasks where participants choose between options that are equal in immediate payoff but differ in their long-term implications or *compensatory value*. For instance, imagine a game where you can either reap a small reward for yourself or help reduce a penalty that will hit later (a sort of repair action). Under LoF/QS, as the end of the game nears, people should

increasingly pick the “repair” option even if both choices give the same immediate points – because the repair option has higher Φ (it fixes something for the end). We’ll observe things like: do people explore fewer options when time is short (do they zero in on certain choices)? Do they gravitate toward actions that prevent future loss or provide relief? How fast do they make decisions under different horizons? We essentially watch how the admissible set $\mathcal{A}(t)$ —the set of actions they consider—changes from wide-open (when there’s plenty of time) to narrow-and-focused (when time is almost up).

- Perturb candidate control hubs: We take advantage of noninvasive brain stimulation techniques like TMS (Transcranial Magnetic Stimulation) or tDCS (transcranial direct current stimulation) to temporarily modulate activity in specific brain regions that we hypothesize contribute to QS-like control dynamics. Chapter 5 identified likely suspects: regions like the right inferior frontal gyrus (rIFG) and anterior cingulate cortex (ACC), which are involved in impulse control and monitoring. So, for a subset of participants, we might apply inhibitory TMS to rIFG (dampening its activity) and apply stimulation over dorsal midline scalp locations intended to modulate dACC-linked networks (with the caveat that standard TMS has limited depth and spatial specificity for deep midline targets), while they perform these horizon-sensitive tasks. The prediction: if those regions are critical for the horizon-based pruning of options, then disrupting them should reduce that effect (people become less discriminating about high- vs. low- Φ actions when time is short), whereas enhancing them should exaggerate the effect. This gives us a causal test – we’re effectively asking, “If we dial down the brain’s control brakes, do people stop doing the last-minute fairness adjustments?”

The aim here is *not* to prove the entire Law of Fairness in one go. Rather, it’s to demand that the hypothesis produce specific fingerprints in behavior and biology on cue. It’s a rigorous challenge: “If this law is real, show us the pattern under these controlled conditions.” If LoF can pass these lab tests, it earns significant credibility.

To bridge theory to experiment, let’s clearly define the key constructs in lab terms:

- Ledger state $L(t)$: This is the cumulative net affect (or “score”) up to time t . In real life, we’d estimate this by integrating our momentary hedonic estimate (e.g., HCl) over time. In the lab, we might simplify it: we could use a running tally of points or a mood rating that updates. Formally, in the wild we described $\hat{L}(t) = \int_0^t HCl(\tau) d\tau$ as the estimated accumulated hedonic state, treating $HCl(\tau)$ as a proxy for momentary felt valence on a common scale. In an experiment, $L(T)$ might just be

“how much good vs. bad you’ve experienced in the task so far” (we can simulate that with rewards and penalties). The participant’s “ledger” can be manipulated: we could start someone off in a deficit (making them work from behind) or in a surplus, to see how that interacts with horizon.

- Horizon H_t : The perceived time or opportunity remaining. In our tasks, this is controlled by instructions or the structure of rounds. We make H_t salient – e.g. a countdown clock or a clear message “Round 8 of 10” so they know how close the end is. We can induce different horizon perceptions experimentally by telling some people the task is almost over when it’s not (a bluff) or by giving some people “extra rounds” unexpectedly. H_t is the lever that we think QS responds to, so we manipulate it and observe.
- Compensability Φ : A quantitative proxy for how much an action can compensate or make up for accumulated losses. We predefine Φ for each choice in our tasks (or, when Φ must be inferred from behavior, we preregister an estimated Φ and use that consistently). For instance, in a decision where Option A yields +10 points to the player and Option B yields +5 points *but* also removes a future penalty or helps a partner (something that compensates for a harm), we’d say Option B has higher Φ than Option A, even though its immediate reward is lower. We might craft a “compensability index” for choices based on features like: does this choice reduce future risk? Does it help recover from past losses or mistakes? Does it benefit someone else or your future self in a way not reflected in immediate points? All those factors contribute to Φ . This allows us to label some choices as high- Φ (more reparative potential) vs. low- Φ (mainly self-indulgent or one-off benefit).
- Admissible set $\mathcal{A}(t)$: The set of options the person is actively considering at time t . We can’t read minds directly, but we infer this from behavior. If out of 5 available options, a participant only ever chooses 2 of them when under time pressure, it suggests the other 3 became “inadmissible” to them under those conditions. We might measure this via the entropy or diversity of their choices. If as H_t shrinks the entropy of chosen actions drops, that implies a narrowing $\mathcal{A}(t)$. We also look for systematic exclusion: e.g. under short horizons, does the participant stop choosing the frivolous options entirely? If yes, those options effectively left the admissible set.
- Shadow price λ_t : This is the theoretical multiplier that scales the value of compensatory actions when time is short. In practice, we estimate λ_t by looking at an interaction effect between compensability and a preregistered monotone

transform of remaining horizon (e.g., a reciprocal-type term defined to avoid division-by-zero at the endpoint). For example, we fit a model to the choices that has a term $\Phi \times H_t^{-1}$ – if that term is positive and significant, it means as H shrinks, high- Φ choices are increasingly favored (the system is effectively “paying extra” for Φ as time runs out). A large λ_t would mean a strong horizon effect; a near-zero λ_t means the person values things the same way regardless of time left.

With these defined, here are the core predictions (lab version) if the QS (Queue System) is indeed operating:

- Menu focusing: When the horizon is short, participants will consider fewer distinct options. Concretely, the diversity (e.g., Shannon entropy) of their chosen actions should decrease as H_t decreases, holding constant trial count and option availability. They don’t necessarily take fewer actions – they might still click 10 times – but those actions come from a narrower subset of possibilities (they “zero in” on what they deem important). For example, someone might normally cycle through many different game strategies (high entropy) when time is ample, but in the final minute, they stick to one or two tried-and-true moves (low entropy). This focusing is not just any narrowing; ideally, it’s a focusing on *high-value* options (next point).
- Compensability tilt: As horizons shrink, choice preferences shift toward options with higher Φ (repair/relief potential) even when *immediate payoffs are held constant*. We expect to see participants pick the “make it right” option more often than the “quick reward” option when they feel the clock ticking. Moreover, reaction times might tell a story: high- Φ choices could become faster (they’re taken more readily, almost instinctively) and low- Φ choices might show hesitation or conflict (if the person considers them at all under time pressure). We might also see persistence: if a participant starts a compensatory action sequence and time is short, they’ll stick with it (keep doing similar high- Φ actions) rather than switching away. Essentially, short horizon = bias towards actions that fix or finalize things, as opposed to actions that just feel good momentarily.
- Control-hub activation: On the neural side, if we record brain data (say EEG or fMRI during these tasks), we predict a specific pattern. When a participant with a short horizon contemplates a low- Φ (indulgent or irrelevant) action, brain regions like rIFG and dACC (which are involved in inhibitory control and conflict detection) should show *boosted activity*. It’s like the brain saying “not a good idea right now” – perhaps manifest as a stronger No-Go P300 or theta burst in ACC indicating suppression of that option. Conversely, the brain’s valuation regions (like

ventromedial prefrontal cortex, vmPFC) should show an enhanced response to high- Φ options when time is short – as if those options suddenly are valued more. And interestingly, the anterior insula and autonomic signals (heart rate, etc.) might show something like a “sigh of relief” response *after* a high- Φ choice is made under pressure. For example, making a compensatory choice might noticeably calm the participant’s arousal (the insula tracks body arousal) when the horizon is short, because the system “knows” a debt was addressed. These are nuanced signals, but we know how to look for them: EEG signatures of cognitive control, fMRI BOLD in decision circuits, changes in skin conductance or HRV after decisions, etc.

- TMS causality effects: If we noninvasively perturb those control regions, the behavior should change in line with our model. Specifically, *down-regulating* the control network (e.g. applying inhibitory TMS to rIFG and dACC) should make participants *less* selective as the horizon shrinks. In other words, the difference between their long-horizon and short-horizon choices will blur – they might keep choosing some low- Φ options even when time is short, because we’ve “removed the brakes” that would normally cut those options out. On the flip side, *up-regulating* those regions (e.g. excitatory stimulation) should *increase* the pruning effect – short-horizon behavior becomes even more focused on compensatory choices than usual. Similarly, perturbing vmPFC should reduce the extra value boost it gives to high- Φ options (so people might not shift as much toward those), and perturbing the insula might affect that post-decision relief signal (so perhaps they don’t feel the relief as clearly and thus might not value the compensatory act as much). These causal tests are sophisticated, but if their outcomes match predictions, it’s a big win for the theory’s credibility.

We also specify what failure patterns in these experiments would look like. For instance, it could turn out that a short horizon just makes people generally more cautious or conservative (a well-known effect in decision-making) rather than specifically targeting compensation. If all we saw was that under time pressure participants avoid risk more or slow down indiscriminately, that wouldn’t count as support for LoF – it would be too nonspecific. We enumerate such alternatives: e.g. if the data show only a generic “do less” or “get anxious” response to short horizons, or if the high- Φ vs. low- Φ choices under short horizon show no difference, then our hypothesis isn’t supported. Essentially, if the $\Phi \times H_t^{-1}$ interaction in our models is null, or if it can be fully explained by something simpler (like “when stressed, people stick to habit”), then the QS account weakens. Our analysis plan includes checks for these. We’ll include control conditions where Φ is the same for all options (to see if horizon alone does anything when compensability isn’t a

factor), and we include tasks where short horizon should not interact with compensability (as preregistered negative controls). These help ensure that if we do see an effect in the main task, it's truly due to the mechanism we think it is.

To implement these experiments, we outline several task families labs can use (details come in subsections). For example:

- “Repair vs. Indulgence” tasks: Participants choose between two actions per trial – one is a “repair” option (fixes something or prevents a future loss) and the other is an “indulgence” option (gives an immediate reward but no long-term benefit). By design, immediate payoffs are balanced. We run this across conditions where the number of trials left is large vs. small, to see how choices shift.
- “Precommitment” tasks: We let participants precommit to a future helpful action when they have a long horizon, and see if they change their mind when that future arrives and horizon is short. LoF would predict they stick to it if it was a compensatory action, even if temptation arises.
- Effort allocation tasks: e.g. participants have a chance to either resolve a “debt” in the game by putting in extra effort or to play around doing something fun, and we manipulate how much time they think they have to do both. With plenty of time, they might do both; with little time, do they prioritize the debt resolution?
- Simulated “life” scenarios: We could use virtual reality or narrative games where participants make decisions for a character over a lifecycle (compressed into an hour). We then introduce a twist like an illness in the story (shrinking the horizon) to see if their decision policy shifts toward wrapping up the character’s story (visiting family, etc.) rather than accumulating more “wealth” in-game. This is more involved, but it directly mirrors life decisions.

Each of these task families will be described, and labs can adopt whichever fits their setup. We want these experiments to be *reproducible*, so we emphasize straightforward designs that many researchers can try, not one esoteric paradigm that only our lab can run.

What you’ll get from this Chapter:

- A concrete plan to simulate fairness dynamics in a lab: You’ll see exactly how we can create scenarios that mimic the pressure of an ending and reveal hidden priorities. This demystifies the idea that “fairness” can only be observed over a lifetime – we show it can be investigated on the scale of minutes with clever task design.

- The key behavioral signals to watch for: We break down the measurable indicators of LoF in action during experiments: reduced choice diversity, shifts toward particular types of decisions, quicker commitments to high-value acts, etc. By the end, you'll know what patterns in data (like graphs of choices over time or reaction time differences) would make us sit up and say "Aha, that's the QS effect!" versus those that would not.
- How specific brain regions come into play: This chapter gives an accessible tour of the neuroscience behind decision control – in particular, why we suspect regions like the rIFG and dACC are the “brakes” and “monitors” enforcing LoF’s constraint, and how vmPFC and insula contribute to value and relief signals. We explain it in intuitive terms (e.g. *“this part of the brain acts like a referee when time is short, calling fouls on wasteful choices”*). You’ll also learn how we can test these roles with tools like TMS, linking mind and brain evidence.
- Distinguishing LoF effects from generic stress or urgency: By walking through potential outcomes, we show how we’ll tell if what we see is truly a fairness-driven effect or just a byproduct of being under pressure. This means you’ll understand the alternate explanations (like “maybe people just get risk-averse when time is almost up”) and how our experiment is designed to rule those out. It’s a lesson in good experimental hygiene: designing controls and negative tests so that a positive result actually means something.
- The rigorous methodology behind the scenes: For those interested, the chapter’s research notes detail how we pre-register these experiments, determine sample sizes to have enough power, choose proper statistical tests (e.g. switching to Negative Binomial regression if needed, as mentioned), and define regions of interest (ROIs) for brain imaging. In everyday language, we outline how we avoid p-hacking and ensure that any discovery is robust. You’ll gain insight into how a seemingly “wild” idea (testing fairness in a lab) is handled with scientific strictness.
- A preview of failure criteria in the lab context: Just as with other chapters, we make clear what lab results would *disconfirm* our theory. You’ll see what it would look like if our horizon experiments showed nothing special – say everyone behaves the same regardless of time limit, or the brain data shows no difference – and why that would force us to rethink. Knowing the failure modes gives you confidence that we’re not bending interpretations; we either see the pattern or we don’t.

Subsections in this Chapter:

- **12.1 Short vs. Long Horizons in the Lab** – Introduces the core experimental approach and one of our primary tasks in detail. We describe a baseline task and then how we create a “short horizon” version of it. Essentially, it sets up how we manipulate the feeling of “time left” in a controlled setting.
- **12.2 Expected Control-Hub Signatures** – Describes what neural and physiological signatures we predict will accompany the behavioral changes. For instance, we talk about expecting stronger inhibitory signals in EEG when someone resists a bad option with little time left, or an uptick in stress markers if they contemplate a low- Φ action under time pressure. We paint a picture of the pattern of brain activity that would validate the QS model (who’s “lighting up” and when). This helps link the psychological theory to concrete neuroscience predictions.
- **12.3 Perturbation: TMS to rIFG/ACC** – Details our plan to use brain stimulation to test causality. We identify the exact targets (right inferior frontal gyrus and dorsal ACC) and rationalize why these are chosen. We explain the protocol: for example, “Participants wear a TMS cap; in one session they receive real inhibitory TMS to rIFG before doing the task, in another they get sham stimulation as a control, comparing their behavior across sessions.” The expected outcomes are reiterated (e.g. with rIFG inhibited, do they choose more low- Φ options than they normally would under time pressure?). We also cover safety and ethics of TMS briefly, assuring that these methods are non-harmful.
- **12.4 Research Notes: Preregistration, Power, ROIs** – Provides technical transparency. We list that we’ve preregistered the experiment design on an open science platform, including our primary and secondary outcomes. We discuss how we calculated needed sample sizes (e.g. based on effect sizes from similar studies on decision-making under time pressure). We mention our use of ROI analysis for fMRI (specifically focusing on rIFG, ACC, vmPFC, insula as defined from prior literature) and how we’ll correct for multiple comparisons if we go fishing beyond those regions. This section might also mention any plan for collaboration or data sharing (since multi-site replication could be key). It’s essentially the fine print that ensures the reader this is being done in the most rigorous way.
- **12.5 Negative Controls** – Here we lay out parallel tests to confirm that any effects we see are indeed due to the mechanism we think. For instance, we might have a version of the task where all options are equal in Φ ; if horizon changes behavior even when there’s no “compensatory” difference, that would suggest a general urgency effect, not LoF specifically. Or we include a condition where we apply TMS to a *control* brain region that shouldn’t affect the behavior (say stimulating the motor cortex or a sham site) to show that only targeting the QS hubs matters. This section describes those control conditions and what outcomes we expect from them (ideally, null results—no effect—if our theory is right, because these

controls are meant to produce nothing). Including this demonstrates how we will catch false positives or misattributions.

- **12.6 Fail Patterns in Lab Tests** – As usual, we conclude the chapter by defining failure clearly. We enumerate things like: “If participants under short horizons show no significant increase in high-Φ choices across N>100 trials, that contradicts prediction P2.” Or “If TMS to rIFG does not significantly change the choice pattern compared to sham in at least 2 out of 3 behavioral metrics, the causal link is not supported.” We set thresholds for what would count as a meaningful signal and acknowledge that not meeting them means LoF did not show up in the lab. We also discuss possible reasons for failure (maybe the lab setup was too artificial, or our manipulations weren’t strong enough) but commit that, if well-powered attempts fail, the burden is on the theory. This section basically draws the line: either we see these lab effects or we reconsider the universality of LoF.

Where we go next:

We’ll begin by walking through a simple experiment that captures the essence of our approach. In Section 12.1, we introduce a concrete task where the only difference between two conditions is how much “future” the participant believes they have. This will make it easy to spot the kinds of behavior shifts – if any – that occur when the horizon abruptly shortens, setting the stage for layering on complexity in later sections.

12.1 Short vs. Long Horizons in the Lab

This section provides ready-to-run protocols for inducing short and long horizons in ordinary behavioral labs without causing undue stress and with minimal (ethics-approved) deception only when required for specific controls. The goal is to alter perceived time remaining—our experimental stand-in for H_t —and measure how the admissible set tilts toward repair/relief options at equal immediate utility. All designs here are intended to be low participant burden and compatible with concurrent recordings (EEG, eye-tracking, HRV), and with subsequent intervention blocks (e.g., introducing TMS after baseline).

12.1.1 Design principles

To ensure clear and interpretable effects, we adhere to a few key design principles:

Equal immediate payoff: In each critical choice, both options should yield the same points or reward *now*. Only their downstream compensability differs. For example, one choice might “help someone later” and the other “is just for fun,” but both give +10 points immediately. This way, any preference shifts aren’t due to short-term gains, only long-term Φ .

Orthogonalize confounds: We match options on every obvious factor besides Φ : effort required, difficulty, sensory stimuli, risk level, and so on are kept as equal as possible. The horizon manipulation (short vs. long) is applied at the block level (not at the individual-trial level) so that, within a block, all trials share the same horizon context. This prevents trial-to-trial strategy swings and focuses the comparison between conditions.

Make horizons salient but not stressful: We want participants to *believe* the horizon condition (e.g. that this really is the final round) without inducing panic or unethical pressure. Techniques include gentle countdowns, “last chance” frames, or limited opportunity windows that *expire*, rather than threatening loss or using deception about real stakes. The idea is to create a sense of closure looming, not to frighten or trick.

Measure the manipulation: We include manipulation check questions after each block. For example, on a 1–7 scale: “I felt the clock was ticking and time mattered in that block,” or “I expected additional chances later.” If participants in the short-horizon condition don’t actually feel the end is near (or vice versa), we need to know that – those data might be analyzed separately or dropped.

Within-subject counterbalancing: Generally, each participant experiences *both* a short-horizon condition and a long-horizon condition, in counterbalanced order. This controls for individual differences in behavior baseline. If there’s a concern about learning or

fatigue across blocks, we can have them on separate days or intersperse filler tasks. The critical point is everyone serves as their own control for the horizon effect.

12.1.2 Core manipulation set

We describe a few concrete ways to implement the horizon difference:

(A) Countdown expiration (minutes-scale horizon): In the short-horizon block, a visible countdown timer (e.g. 5:00 → 0:00) is running on screen. Participants are told that certain meaningful actions (like sending an apology to an online partner, fixing an error in a document, freeing a “stuck” helper bot) are only available while time remains. When the timer hits zero, those high-Φ actions disappear, leaving only neutral options. In the long-horizon block, participants complete an identical number of trials but with no visible timer and an instruction that “there will be additional opportunities later” to use those special actions. Checks: After each block, we ask if they felt time pressure or finality (e.g., “I had time to make things right” – they should disagree for short horizon, agree for long).

(B) Last-chance framing (hours-to-days scale horizon): The scenario is framed as helping a partner over multiple sessions. In the short-horizon condition, participants are told “*This is the final session – your last opportunity to improve your partner’s outcome.*” In the long-horizon condition: “*This is session 1 of several – there will be more chances later.*” The choices within the session are the same in both cases. This tests whether merely believing “it’s the last chance” versus “I can always fix things later” alters behavior. Checks: We probe perceived finality (“I felt this was my only shot to help”), and expectation of future chances.

(C) Opportunity depletion (trial-count horizon): Participants have a limited number of “repair tickets” to use across trials. In the short-horizon version, they might get, say, 3 tickets total. Using one on a trial does something beneficial (like improving someone’s later payoff), but when tickets are gone, they cannot do that anymore. In the long-horizon version, the tickets replenish over time or there’s an ample supply, reducing urgency. Here the horizon is expressed in terms of a depleting resource of opportunities. Checks: We ask if they felt the repair opportunities were scarce or if they could “save some for later,” etc., to confirm the manipulation.

All these paradigms create an environment where participants face the *same immediate trade-offs* but under different temporal mindsets. By comparing short vs. long horizon blocks, we isolate how horizon perception influences the inclination toward compensatory acts.

Expected behavioral outcomes: In these tasks, if the Law of Fairness dynamics are present, we anticipate that in short-horizon conditions participants will (i) choose the

high-Φ option more often, even though it has no immediate advantage, (ii) execute those choices faster (a sign of less ambivalence), and (iii) be more likely to invest effort into reparative sequences (or conversely, quit low-value sequences sooner). In long-horizon conditions, we expect more “business-as-usual” behavior: participants might occasionally choose the indulgent or easier path since they subconsciously feel there’s time to course-correct later.

If instead we observe no difference – e.g. participants treat a “last session” the same as an ongoing one – that would be an important null result suggesting horizon by itself isn’t affecting choices (or our lab simulation of it isn’t strong enough). Section 12.6 will revisit these possible Fail patterns and what to conclude from them.

12.1.3 Where we go next:

With the basic task designs in hand, we move next to what we should see in the brain and body if the Queue System is indeed working during these tasks. Section 12.2 details the neural and physiological signatures to look for as hallmarks of horizon-sensitive control.

12.2 Expected Control-Hub Signatures

If the Queue System (QS) is real, it should leave distinctive, replicable fingerprints in the brain-and-body control network when horizons shrink and compensability (Φ) becomes decisive. This section specifies *what* to look for, *where* in the brain/body, and *when* in time – first in plain language, then in terms of concrete signals one could preregister. Each signature comes with a predicted direction of effect, the timing (when it should occur relative to events), factors that might modulate its strength, and clear falsification criteria.

12.2.1 The “Control Quartet”: rIFG, dACC, vmPFC/OFC, anterior insula

We focus on four key regions as the hypothesized control network implementing QS dynamics:

Right inferior frontal gyrus (rIFG): Often linked to inhibitory control or “braking” of actions. Prediction: When the horizon is short, *and* a low- Φ option is on the table, rIFG should show a stronger, earlier burst of inhibitory activity, consistent with increased inhibitory control engagement for options with lower compensability under time pressure. In EEG, this might appear as increased beta-band power (15–30 Hz) over right frontal leads during decision onset; in fMRI, higher BOLD in rIFG on trials where a low- Φ option was *available* but perhaps not chosen.

Dorsal anterior cingulate cortex (dACC): A region for conflict monitoring and effortful control allocation. Prediction: dACC will ramp up its activity more steeply for low- Φ choices under short horizons. For instance, if a participant embarks on a low-value sequence with little time left, dACC should show rising theta-band activity (4–7 Hz) and BOLD signal, reflecting internal conflict or cost of continuing.

Ventromedial prefrontal cortex / orbitofrontal cortex (vmPFC/OFC): A value integration hub. Prediction: At equal immediate utility, vmPFC will assign a *higher subjective value* to high- Φ options, especially when the horizon is short. In other words, a choice that offers future relief will elicit an outsized value signal in vmPFC when time is running out – as if the brain applies a “compensatory bonus.” This is effectively a neural correlate of the shadow price λ_t .

Anterior insula (aINS) and associated autonomies: Involved in interoception and negative arousal. Prediction: After a participant chooses a high- Φ (compensatory) option in a short-horizon context, the anterior insula should *quiet down* more quickly and autonomic arousal should normalize faster. This would manifest as, say, a quicker recovery of heart rate variability (HRV goes up as stress is relieved), skin conductance

response decays faster, and pupil size returns to baseline – indicating the person’s internal state “settled” because they did something to address a looming deficit.

Together, these four should orchestrate QS behavior: rIFG and dACC prune and push back against low-value actions, vmPFC amplifies the appeal of high-value (repair) actions, and insula/autonomic systems register when a compensatory action has alleviated the “pressure,” returning toward baseline. If all goes as predicted, we’d see a coordinated pattern across them as horizons tighten.

12.2.2 Behavioral preconditions for neural signals

Before attributing meaning to any neural or physiological differences, we set a ground rule: the neural/physio signatures are only considered valid if the behavioral interaction is present. That is, we must see the basic effect in choices (from Section 12.1) – more high- Φ selections and faster response times under short horizon, etc. If the behavior shows no $\Phi \times (\text{Horizon})$ interaction, any differences in brain signals would be hard to interpret and likely moot. In practice, this means we use the behavioral results as a filter: neural divergences that don’t accompany any change in behavior would not count in favor of QS (they might just be noise or unrelated effects).

So, assuming we *do* have the behavioral evidence in hand, we proceed to the specific neural markers.

12.2.3 rIFG – the selective brake

Prediction: In short-horizon blocks, low- Φ options (those indulgent or non-reparative choices) are predicted to be associated with stronger inhibitory bursts in the 150–300 ms range after the option appears. This could be seen as increased beta-band power in EEG over right frontal sites (beta is often associated with inhibitory control). Additionally, just before action execution (e.g. in the split second before a button press), we expect heightened rIFG engagement if the action being taken is low- Φ under a short horizon – as if rIFG is trying to stop it. In fMRI, we’d predict overall higher activation of rIFG during short-horizon blocks, particularly on trials where low- Φ temptations existed.

Timing: EEG/MEG might show a burst of right-frontal beta ~200 ms after the stimulus onset (when evaluating the options), and possibly a sustained beta or even outright stopping-signal-like activity ~250 ms before a movement on “hard to justify” actions. Think of it like a blip of “Are you sure you want to do that?”.

Modulators: The effect should be larger when the “stakes” are clear – e.g. when compensatory channels are open (so a low- Φ choice truly is a lost opportunity to help)

and when the participant's current ledger \tilde{L} is in bad shape (they *need* to do something good). Those conditions would heighten rIFG's response.

Falsifier: If we observe that rIFG activity doesn't discriminate between low- and high- Φ options in short horizons – or if any increased activity is fully explainable by some generic factor like the option being more cognitively difficult or risky – then the QS-specific claim for rIFG weakens. For example, if rIFG lights up equally for all choices or only correlates with, say, the number of rules or the presence of any conflict (not specifically the Φ aspect), then it's not acting as a selective brake as hypothesized.

Secondary markers: We could also look at EMG (muscle activity) on participants' hands for partial movements that get inhibited. We might see more instances of a muscle twitch that doesn't complete (a sign of a motor plan that was aborted) on short-horizon low- Φ trials. That would be a neat peripheral indicator of rIFG at work.

12.2.4 dACC – the cost of continued pursuit

Prediction: ACC is predicted to show stronger theta-band activity (4–7 Hz oscillations) and higher BOLD signals specifically for low- Φ choices, especially if the person is trying to carry them out in a short horizon. If a participant persists in a low-compensatory action when time is almost up, we expect dACC to increasingly signal “this is costly/inefficient.” We also predict a *stepwise increase* in ACC activity with each step of a low- Φ sequence under short horizon – reflecting growing conflict or a mounting “urge to stop.”

Timing: We might see a theta power increase ~250–600 ms after option onset (when deciding to continue a low-value action), and a ramping pattern across time if the action has multiple stages. In an fMRI block where someone is performing continuous actions, dACC might show an escalation that peaks right before they finally quit the task.

Modulators: The effect is stronger when the horizon cues are explicit (the participant is vividly aware time is almost up) and when the low- Φ sequence is something that *cannot* be easily reversed or compensated later (high stakes). If tasks are trivial or easily redone, ACC might not bother ramping.

Falsifier: If ACC's activity can be explained entirely by other factors – e.g., it only tracks how physically hard the task is or simply how much time on task has passed, without any interaction with the Φ aspect – then our interpretation is wrong. For instance, if ACC theta increases for long tasks irrespective of compensability, or if it fires for any high-conflict choice even when horizons are long, then it's not specifically reflecting the QS dynamic.

Cross-check: One could check if ACC theta actually predicts *when* a person aborts a low- Φ sequence. We'd expect higher ACC signals just before they give up (which would be consistent with QS: the system finally forces a stop).

12.2.5 vmPFC/OFC – value with a compensatory bonus

Prediction: When immediate payoffs are equal, vmPFC will show a value signal bias favoring high- Φ options, and this bias will grow stronger as the horizon shortens. Effectively, the brain's valuation center acts as if the high-compensability choice is worth more when there's not much time left, consistent with a model in which an additional compensability term is weighted more heavily as horizon shortens. In modeling terms, we'd see a significant interaction such that vmPFC encodes value = (immediate reward + $\lambda \times \Phi$), with λ_t larger as H_t decreases (using the preregistered monotone transform of horizon).

Timing: In EEG terms, something like the *frontal P3* or other reward expectancy signals ~300–700 ms post-stimulus might be larger for the high- Φ item under short horizon. In fMRI, vmPFC BOLD might be higher on short-horizon trials where the participant chooses (or even just considers) the compensatory option, with peak activation a few seconds after the choice.

Modulators: This effect should be *bigger* if the person currently has a “debt” in their ledger (they're in a negative state and could really use a good deed) and if the opportunity for relief/repair is genuinely available. If we experimentally *close* the channel (no real impact of the choice), the vmPFC might not show the bonus because subconsciously the person may detect that it's futile.

Falsifier: If, after controlling for obvious factors (like risk, effort, or framing effects), we find no value difference in vmPFC between the high- and low- Φ options, or worse, if we find the opposite (e.g. people value indulgence more under short horizon), then the theory's value prediction fails. Also, if any vmPFC difference we see can be explained away by immediate reward differences or other non- Φ attributes, then it's not evidence for a compensatory bonus.

Cross-check: A fancy approach is MVPA (multivariate pattern analysis): train a classifier on vmPFC patterns for normal reward vs. non-reward trials. If QS is real, that classifier might under-predict the neural activation for a high- Φ choice in a short-horizon context – indicating that the vmPFC response is larger than expected from immediate reward alone (hinting at an extra ingredient, presumably Φ).

12.2.6 Anterior insula and autonomic settling

Prediction: Choosing a high- Φ action when the horizon is short is predicted to be associated with a kind of “sigh of relief” in the body and insula. After the choice, the anterior insula’s activity (and associated arousal responses) should decline more quickly compared to other choices. Essentially, if you do the thing that helps “balance the books” when time is almost up, your body should register that as a stress reduction.

Timing: We’d measure things like: the peak pupil dilation or skin conductance after making a choice, and how long it takes to return to baseline. For high- Φ choices in short-horizon blocks, we expect any arousal spike to be short-lived – maybe within 1–3 seconds the pupil is back down, EDA is dropping, HRV (a measure of parasympathetic activation) is improving. In fMRI, the insula might show a brief activation at choice moment but then drop, whereas with a low- Φ choice it might stay elevated longer.

Modulators: If the compensatory choice truly addressed a pressing need (like alleviated pain or prevented a loss), the settling effect is stronger. It’s also stronger if the participant is someone who feels anxiety relief strongly when they fix things (individual difference). If the channels were closed (no actual effect of their choice), we might not see much settling because nothing was truly resolved.

Falsifier: If high- Φ and low- Φ choices produce identical physiological trajectories once you control for basic factors like effort duration or movement, then there’s no evidence of a special settling effect. Also, if any differences appear even when nothing reparative was done (e.g. they choose a random option and still show a drop in arousal), then our interpretation is off – maybe it was just relief that the task ended, etc., not specific to Φ .

12.2.7 Network coordination – who leads, who follows

It’s one thing to look at each region, but QS suggests an integrated control process. We can consider effective connectivity:

Prediction: In short-horizon conditions, the causal flow between regions might shift. For example, we expect more top-down inhibition signals: rIFG might exert greater influence on motor regions (suppressing actions) and ACC might drive rIFG more strongly (signaling “apply the brake!”) on low- Φ trials. Meanwhile, for high- Φ decisions, vmPFC (valuation) might more strongly drive ACC and rIFG to stand down (since it’s a “good” choice, less need for braking).

Tools: Analyses like Granger causality or dynamic causal modeling on EEG/MEG, or psychophysiological interaction (PPI) analyses on fMRI data, can test whether preregistered effective-connectivity proxies change with condition (interpreted

cautiously, since such methods are model-dependent and do not by themselves establish biological causation). For instance, do we see increased coupling from ACC→rIFG in short horizon during tough choices? Or vmPFC→ACC when a compensatory choice is being evaluated?

Falsifier: If we find that any connectivity changes are fully explained by generic difficulty or timing (e.g. maybe in short blocks everything is just more hurried so all connections ramp up similarly) and show *no dependence* on Φ or channel status, then we can't credit it to QS. We're specifically looking for $\Phi \times \text{Horizon}$ interaction effects in connectivity: certain pathways strengthening only in the scenario where they're supposed to (low- Φ /short horizon).

12.2.8 The omnibus test: $\Phi \times \text{horizon}^{-1}$ everywhere

To tie the above together: across all these modalities, the interaction between compensability and horizon is the unifying signature:

In rIFG/ACC: significantly larger inhibitory/conflict signals for low- Φ *when horizon is short* (vs. *long*).

In vmPFC: a larger value surplus for high- Φ *when horizon is short*.

In insula/autonomics: faster settling after high- Φ *when horizon is short*.

This $\Phi \times \text{Horizon}$ interaction is our "omnibus" criterion. If all modalities show it, that's a powerful confirmation.

Falsifier (omnibus): After properly accounting for nuisances (effort, risk, general arousal, etc.), if *all three* of those interactions turn out null or even opposite – or if what we observe can be entirely captured by, say, a risk aversion variable or a fatigue effect – then the QS hypothesis does not hold in this context. We would have essentially a multi-channel null.

12.2.9 The Role of Channel Availability (Moderation)

A crucial aspect is testing whether these effects *require that compensatory actions are actually available*:

Prediction: If we open relief/repair channels (e.g. the participant truly can help someone or fix something), all the QS signatures should be amplified. If we close those channels (making compensatory moves ineffectual or not present), the signatures should be attenuated, even if the horizon is short. In behavior: the tilt toward high- Φ should appear only when those high- Φ options genuinely mean something. In neural data: vmPFC, rIFG, etc., should react strongly only when a real affordance is at stake.

Falsifier: If our results show no change when channels are closed versus open – or paradoxically, stronger effects when nothing real can be done – then we’re likely not measuring what we think we are. For example, if even a “fake” repair option (one that doesn’t actually do anything) triggers the whole pattern, then perhaps the effect is just driven by framing or demand characteristics rather than the hypothesized mechanism.

12.2.10 Individual differences that (hopefully) don’t erase the pattern

We acknowledge that people differ, but the core QS effect should transcend certain individual traits:

Traits: Personality factors like agreeableness or conscientiousness might make someone more likely overall to help (higher baseline Φ choices), but they *should not eliminate the horizon interaction*. A highly altruistic person might choose high- Φ options even with a long horizon, but if QS is universal, they would *still* ramp up even more when the horizon shortens.

Mood/clinical states: Someone with mild depression or anxiety might have higher baseline arousal or a different starting point, but we expect the short-horizon vs. long-horizon pattern to still occur once we control for that baseline.

Age: Older adults might react more slowly or have smaller autonomic ranges, but again the *direction* of the effect (favoring compensatory acts as horizons shrink) should remain the same, even if magnitudes vary.

Falsifier: If adding a simple covariate like age or a personality score to the analysis makes the $\Phi \times$ Horizon interaction vanish, that would imply what we observed was specific to a subgroup and not a general effect. E.g., if only highly agreeable people show it and others don’t at all, then maybe it’s not a fundamental process but a personality-driven one.

12.2.11 Minimal reporting set (ensuring comparability)

To make sure different studies can be compared or aggregated, we propose a minimal set of results everyone reports:

Behavior: The odds ratio (or equivalent effect size) for choosing high- Φ in short vs. long horizon, with confidence intervals. Also the difference in mean reaction time for high- Φ vs. low- Φ choices in short horizon, and any change in entropy of choices.

Neural: Effect sizes (Cohen’s d or percent signal change) for the key interactions in rIFG (beta power or BOLD), ACC (theta power or BOLD), vmPFC (BOLD), etc. And share the exact ROI definitions and any preprocessing steps used (so others can do the same).

Autonomic: Report the change in HRV (e.g. Δ RMSSD) post-choice between conditions, the difference in EDA recovery time, and pupil dilation differences for key conditions.

Manipulation checks: The average scores on the horizon perception questions, to document that participants indeed felt the difference, and how many participants might have been excluded for not believing the scenario.

Adversarial model fits: If we compared our QS model to alternative models (risk-only, etc.), report those fit indices (AIC, BIC, or cross-validated log-loss differences) so we know how much better or not QS explained things.

This consistency will help meta-analyses later. It also forces completeness: for instance, even if a hypothesis “failed,” reporting the negative control outcomes (like maybe the risk-only task indeed showed null, good) is important.

12.2.12 What success vs. failure looks like (summary)

To recap in plain terms:

A supportive outcome would be: behavior clearly shows the $\Phi \times$ Horizon interaction (people skew toward making things right when time is short), *and* the brain shows rIFG and ACC specifically reacting to low- Φ under short horizons, vmPFC giving extra weight to high- Φ , and autonomic signals calming after high- Φ actions – especially when real opportunities to compensate are present. That would strongly support that a QS-like mechanism is in play.

An ambiguous or partial outcome might be: we see some of the effect but not all. For example, maybe behavior and ACC signals show it, but rIFG doesn’t or vmPFC doesn’t. In that case, we’d treat it as “under strain” – perhaps we need to refine the experiment (maybe the horizon manipulation wasn’t strong enough or the brain measure was too noisy). We wouldn’t throw out the theory on one partial result, but we also wouldn’t claim victory.

A disconfirming outcome would be: after doing everything with good power, *nothing* shows the interaction as predicted. For instance, people’s choices don’t change with horizon, and simultaneously rIFG/ACC don’t care about Φ any more than usual, etc. Or maybe we find that a simpler explanation (like “they were just more risk-averse with less time”) accounts for what we see across the board. Such a pattern, especially if replicated, would seriously challenge the QS account.

So far, everything we’ve discussed is *correlational* or observational – we manipulate conditions and observe behavior and brain responses. Correlation, however, is not causation. In Section 12.3, we take the critical next step: we *poke the system* by

perturbing those control hubs (rIFG and ACC) directly. If QS is genuinely the mechanism driving these patterns, then tuning those brain regions up or down should causally alter the horizon effect. Dialing down the “brakes” should make people behave as if the horizon doesn’t matter (flattening the difference between last chances and abundant chances); dialing up the brakes should exaggerate the difference. The next section provides a turn-key protocol for these causal tests and what to look for in the outcomes.

12.3 Perturbation: TMS to rIFG/ACC

Correlation maps the playing field; perturbation tests whether the field governs the play. Now we move to *causal* experiments: using transcranial magnetic stimulation (TMS) to modulate activity in the suspected control hubs (right inferior frontal gyrus and dorsal ACC) and observing the effect on horizon-dependent behavior. If the Queue System tightens and tilts the admissible set as horizons shrink, then inhibiting those hubs should flatten that effect (people will be less able to preferentially prune low- Φ actions), while enhancing those hubs should steepen it (they become even more strict about what actions “pass”). This section lays out a protocol for such noninvasive causal tests.

12.3.1 Core hypotheses for perturbation

We formalize three main hypotheses:

H1: rIFG as causal brake. *Down-modulating* rIFG (e.g. using continuous theta-burst stimulation, cTBS, which tends to inhibit) will lead to more low- Φ choices in short-horizon blocks, reduce the usual “inhibitory” signatures (like fewer beta bursts or less EMG suppression), and overall flatten the $\Phi \times$ Horizon interaction in behavior. Conversely, *up-modulating* rIFG (intermittent TBS, which can excite) should do the opposite: fewer low- Φ choices and an even stronger interaction (short horizons become extremely selective). In short, if rIFG is the brake, taking your foot off it (down-modulate) should let low-value actions slip through even when time is short.

H2: ACC as control allocator. *Down-modulating* dACC will reduce mid-sequence “stalling” on low- Φ policies and blunt those theta ramps we expect before aborts. Essentially, if ACC isn’t fully online, participants might continue low-value actions longer than they should (they lose the signal to stop) – making short-horizon blocks look behaviorally more like long-horizon. *Up-modulating* dACC should heighten sensitivity: people might bail out of low- Φ sequences even earlier and with more pronounced signals.

H3: Specificity and channel dependence. The above effects should be specific to the conditions we think matter. They should be *largest* when repair/relief channels are open (because only then does QS have something to actually do) and minimal or none in

control conditions. Also, if we run a control task that has similar difficulty or timing but no compensability element (e.g. a pure risk task), TMS should *not* systematically shift those choices – any effect should really hinge on QS being engaged.

These hypotheses will be tested by comparing behavior and neural measures across Sham vs. Active TMS conditions in the short vs. long horizon tasks.

12.3.2 Targeting and neuronavigation

Precise targeting of rIFG and dACC is crucial:

rIFG (right inferior frontal gyrus): We target the pars opercularis/triangularis region of the right IFG. A rough MNI coordinate to aim for is around ($x = +52, y = +14, z = +18$), used here as an illustrative starting point for neuronavigation rather than a claim of a single definitive locus. To refine targeting, we can use each participant's MRI and even have them do a quick Stop-Signal task in an fMRI or EEG to localize where their rIFG "lights up" during inhibition – then use that as the spot.

dACC (dorsal anterior cingulate / mid-cingulate): This is deeper and along the midline. We use coordinates around ($x = +4, y = +24, z = +36$) as a starting point, but interpret coil-based stimulation here as an attempt to modulate dorsal medial frontal circuitry linked to dACC rather than a focal stimulation of dACC itself. Because it's midline, we place the TMS coil slightly off-center on the scalp (just to one side of the midline, since figure-eight coils have a focal point under the intersection). We might have participants do a Stroop task in fMRI or look at midline theta in EEG to get a personal hotspot for ACC conflict signal.

Coil orientation: For rIFG, a typical orientation is with the coil handle pointing posteriorly and laterally (to induce a current roughly perpendicular to the gyrus orientation). For dACC, we might orient the coil front-to-back over the top of the head (since dACC is below the vertex) and use a slightly angled approach to maximize reaching it. These details matter to consistently engage the targets.

Neuronavigation: We employ an optical tracking system with each participant's head MRI to ensure the coil is placed over the specified coordinate and held at the correct angle, usually keeping within ~3 mm and a few degrees of the target throughout the session. This is standard in research TMS to maintain precision.

All these steps help us actually stimulate *the same functional areas* across participants, rather than haphazardly zapping.

12.3.3 Stimulation paradigms and safety

We choose TMS protocols that can create a lasting modulation for the duration of a task block:

Specifically, cTBS (continuous theta-burst) to *inhibit* and iTBS (intermittent theta-burst) to *excite*. cTBS consists of bursts of 3 pulses at 50 Hz repeated at a rate of 5 Hz (i.e. every 200 ms) for about 40 s (total ~600 pulses). iTBS uses the same bursts but delivered in 2 s trains with 8 s pauses over a few minutes (also totaling ~600 pulses). Both are delivered at ~80% of each participant's motor threshold (as defined by standard motor-threshold procedures). These protocols often produce after-effects on the order of tens of minutes, but both direction and duration can vary substantially across individuals and targets; therefore, stimulation effects must be verified behaviorally within-session rather than assumed from protocol alone.

Sham: For a control condition, we either use a sham coil that makes the same clicking noise but doesn't deliver a field effectively, or we angle the real coil 90° away from the scalp (so the induced field in cortex is greatly attenuated) while providing skin sensations (like electrodes on the scalp that mimic the feeling). This way participants experience the same setup and sound but presumably no significant neural effect.

Timing: We apply the TBS before the task block. For example, a schedule might be: do a quick localizer task (to engage the region momentarily), then immediately deliver cTBS (~40 s) or iTBS (a few minutes), then within the next minute start the short-horizon task which runs ~10–15 minutes, hopefully within the window of the TMS aftereffect. Aftereffects are often on the order of tens of minutes (variable across individuals and targets), which typically covers a single task block.

Screening and safety: Every participant is thoroughly screened for TMS contraindications (e.g., personal history of seizures/epilepsy, relevant implants/foreign bodies, and other standard contraindications assessed by a preregistered screening protocol). We measure their motor threshold to calibrate intensity. Earplugs are given for the loud clicks. And we have emergency procedures in place (like what to do in the unlikely event of a seizure – which is exceedingly rare at these parameters). We also keep coil positioning such that it minimizes any muscle twitching (for comfort) and monitor them throughout.

Ethically, we also debrief participants afterward about what we were doing (while still maintaining blinding for any future sessions they might have). The key is we want to alter their cognitive state slightly (e.g. slower brake or faster brake) without causing discomfort or risk.

12.3.4 Experimental design

We touched on it, but let's lay out the design structure clearly:

Within-subject design: Each participant goes through multiple sessions (for example, three sessions): one with Sham stimulation, one with Active rIFG stimulation, and one with Active dACC stimulation. The order of these sessions is randomized across participants to avoid order effects.

Session structure: In each session, after setup and baseline measures, we might do a brief localizer block (~10 minutes) – e.g. a Stop Signal task or Stroop, to engage the target region and also to measure its baseline function. Then we apply the stimulation (cTBS or iTBS as appropriate). Immediately after, the participant performs the main horizon task in two blocks: one short-horizon, one long-horizon (with block order counterbalanced across sessions). We might include a short rest between blocks. The idea is to assess the interactive effect: Horizon (short vs. long) × Stimulation condition (rIFG down vs. sham vs. etc.).

Sample size: Because within-subject TMS effects and individual responsiveness can be variable, sample sizes should be set by preregistered power/simulation analyses tailored to the specific task, outcome model, and expected effect sizes, with explicit allowance for dropouts and session exclusions.

Counterbalancing: We ensure equal representation of all order permutations (e.g. a third start with Sham, a third with rIFG, a third with ACC; also half do short horizon first vs. long first in each session, etc.). This prevents confounds like “people always do better on the second session because they learned the task” from aligning with any one stimulation condition.

12.3.5 Primary endpoints (behavioral)

What effects do we expect to actually measure from these perturbations? The primary behavioral outcomes:

Choice preference: We will look at the change in the odds of choosing high-Φ options under short vs. long horizons, comparing when TMS is active vs. sham. For example, under Sham, say the odds ratio of high-Φ (short vs. long) is 2.0 (just hypothetical). Under rIFG down-mod, maybe it drops to 1.1 (nearly flat), whereas under rIFG up-mod it might go to 3.0 (even stronger tilt). We'll quantify this as an interaction effect or a difference in differences.

Prediction: Down-modulating rIFG or dACC reduces the $\Phi \times$ Horizon interaction (flatter pruning, people act almost the same in short vs. long). Up-modulating them increases the interaction (bigger gap between short and long horizon behavior).

Reaction time and persistence: We expect, for instance, that normally in short horizon people respond faster for high- Φ . If we knock down rIFG/ACC, that RT advantage might shrink (they become more indecisive or slower, perhaps because the “urgency” signal was disrupted). Also, normally people hesitate or abort on low- Φ sequences under time pressure; with ACC down, they might *not* hesitate as much and might carry on longer than they should. So, metrics like “proportion of low- Φ sequences aborted” in short horizon could increase with up-modulation (people abort even more readily) and decrease with down-modulation.

Task performance consequences: We might also measure the actual accumulated reward or “ledger improvement” achieved. If QS is helpful, then down-modulating it might cause participants to end with a worse ledger (because they failed to prioritize fixes), whereas up-mod could lead to better final outcomes. This is more exploratory but could be interesting.

12.3.6 Primary endpoints (neural/physiological signals)

If we have EEG or other recordings during TMS (or immediately after, or interleaved in a separate run), we can check a few neural predictions as well:

rIFG session: With rIFG down-modulated, we expect to see a decrease in those beta-burst signatures before stopping an action. For example, beta bursts on short-horizon low- Φ trials should decrease after cTBS to rIFG and increase after iTBS. Also EMG might show more unchecked partial movements when rIFG is inhibited (the brake is weaker, so more “flinches” get through).

dACC session: With dACC down, the midline theta ramp we expected before aborting a low- Φ sequence should flatten. Up-mod should exaggerate it (maybe a stronger theta peak). Also, behaviorally, down-mod ACC might lead to fewer abandoned sequences (because the signal to quit doesn’t build up as strongly).

Cross-check in signals: We will also watch vmPFC’s value signal for high- Φ . We *don’t* expect that to vanish unless our TMS spread unintentionally. For example, if when we hit rIFG we also accidentally affected nearby vmPFC networks, one clue would be that the value difference collapsed, suggesting a non-specific effect (arousal or slight discomfort affecting everything). But ideally, rIFG TMS should leave vmPFC signals intact (it only affects the braking, not the value assignment).

Many labs won't run concurrent TMS-fMRI (since it's technically challenging), but we might do separate sessions or pre-post scans. The main thing is: Does perturbing these hubs actually change the neural implementation of QS as we think? We'd like to see that in the EEG/fNIRS etc., not just behavior, to be sure we are hitting the target mechanism.

12.3.7 Analysis model (pre-registered)

To analyze the combined effects, we plan a mixed-effects regression model. For example, a logistic generalized linear mixed model (GLMM) for choices:

$\text{logit}(P(\text{choose high-}\Phi)) \sim a + b_{\text{ID}} + H_{\text{cond}} + \Phi \times \text{Stim} + (H_{\text{cond}} \times \Phi \times \text{Stim}) + (\Phi \times \text{Stim}) + (H_{\text{cond}} \times \Phi \times \text{Stim}) + C$, where H_{cond} is the horizon condition (short vs. long), Stim is the stimulation condition (active vs. sham, or potentially two dummy variables for rIFG vs. dACC vs. sham), and C represents covariates. We include a random intercept b_{ID} for each participant. The primary test is the three-way interaction $H_{\text{cond}} \times \Phi \times \text{Stim}$.

For example, we test whether the difference between short and long horizons in the proportion of high- Φ choices is significantly different under active TMS compared to sham. This is the key effect of interest.

We would run analogous models for reaction time (maybe using a linear mixed model or transforming RT) and for sequence abort rates, etc.

For neural data, we might do an ANOVA or mixed model on the neural metrics (like beta power or theta power) with factors Horizon, Phi (or choice type), and Stim.

We will include nuisance covariates C such as trial order, any trend of fatigue, immediate utility differences (should be none by design, but just in case), individual trait scores, and importantly the coil-to-target distance or stimulation intensity differences. These help soak up variance and ensure any null isn't due to, say, coil being a bit farther on some people.

We also verify that the sham condition shows the normal pattern (to ensure participants in the TMS study behaved as expected when not stimulated).

12.3.8 Specificity and channel moderation

Within each stimulation session, we can also examine if the TMS effects themselves depend on whether channels are open or closed:

For example, if in some blocks we secretly removed the compensatory options (closed channel), does TMS still have an effect? Prediction: The influence of TMS should be larger

when channels are open and minimal when closed. Because if nothing to repair, turning up or down the “brake” might not visibly change anything.

Falsifier: If we find equal TMS effects even in a scenario where no compensatory actions were possible (or worse, a bigger effect when closed), that indicates we might be affecting something more general like impulsivity or attention, not specifically QS.

This is a check on our causal interpretation: we want to see the changes primarily in contexts relevant to QS.

12.3.9 Negative and orthogonal controls

We include a couple of extra control conditions to ensure the TMS isn’t doing something broad or weird:

Site control: We might have a condition where we place the TMS coil on a control site like the vertex (top of head) or left angular gyrus – somewhere not implicated in our process. If we run a few participants with TMS at vertex, we expect no change in the three-way interaction ($\text{horizon} \times \Phi$) compared to sham. This helps confirm that any effect we got from rIFG or dACC wasn’t just due to general brain stimulation or arousal.

Task control: We might include a block of a risk-only task or another decision task that has similar difficulty but no Φ difference. Under normal circumstances, horizon doesn’t matter there. With TMS, it still shouldn’t – if we found, say, that rIFG TMS made people more risk-taking, that would be an off-target effect we need to account for.

Physio control: For EEG, we’ll look at things like blink rate or eye movement artifacts to ensure our results aren’t due to TMS making people blink more or something silly. If necessary we include those as regressors. We expect the QS effects to remain even after removing such nuisance factors.

Overall, these controls strengthen any causal claim by showing it’s specific to the intended network and decision context.

12.3.10 Potential confounds and fixes

Even with careful design, TMS studies have typical confounds; we address the main ones:

Spread to adjacent regions (e.g. premotor or M1): TMS might not only affect our target but also neighboring areas. For rIFG, a risk is stimulating motor cortex if placed too far back. For dACC, risk is hitting supplementary motor area. Fix: We monitor motor evoked potentials (MEPs) in a hand muscle if we suspect motor cortex involvement; also record coil-to-cortex distance and the induced electric field strength in the target vs.

surrounding areas. We can then exclude any session where clearly the wrong area was primarily hit or statistically control for it.

Arousal/startle effects: TMS has an auditory click and somatosensory scalp sensation that can jolt participants, potentially affecting pupil size etc. Fix: Our sham condition controls for the sound, but we also ensure the real and sham conditions are equally loud. We can measure pupil dilation at TMS onset and include that as a covariate if needed. If the effects we see vanish after controlling for pupil jump, then maybe it was just an arousal difference – which would be a problem for interpretation.

Expectation/placebo effects: If participants realize “hey, this session I got real TMS and I feel a bit different,” their behavior could change just from that belief. Fix: Double-blinding as much as possible. Also, after each session, we ask them to guess if it was real or sham. We include that guess as a covariate. If someone was sure they got stimulation and behaved a certain way, we’ll see it. Ideally, they shouldn’t be able to tell reliably.

By preemptively tackling these, we hope any result stands up to scrutiny.

12.3.11 Interpreting outcomes of perturbation

When we get the data, we’ll classify the outcome as:

Strong support: If we see a significant three-way interaction in behavior (the horizon effect is clearly different under active TMS vs. sham) *and* we see the corresponding changes in the brain signals (e.g. rIFG TMS reduced beta bursts exactly as predicted) *and* none of the control conditions show false effects. Moreover, if channel-open vs. channel-closed shows the moderation expected. This scenario would indicate a successful causal validation of QS.

Partial support: Maybe we get the behavioral change (say, rIFG TMS did flatten the horizon effect a bit), but the neural signals were too noisy to detect, or only one of the two target regions showed the effect while the other didn’t reach significance. In this case, the door is open but we’d likely need a replication with higher signal-to-noise (maybe using combined TMS-fNIRS or more subjects).

Disconfirmation: If no difference is seen between active TMS and sham in the critical interaction (e.g. the short-vs.-long difference in choices remains identical), *or* if we see equal changes even in the control site condition, then the perturbation failed to support the QS mechanism. Another bad outcome would be if stimulating our regions causes a broad collapse of behavior (e.g. they become randomly erratic or vmPFC value signals drop out, suggesting we just mucked up the brain nonspecifically). In those cases, the

conclusion might be that either these regions aren't as important as thought or our method didn't properly target the mechanism.

If we hit a disconfirmation, the advice (per our playbook) might be to revisit the model: perhaps QS is not implemented where we thought, or perhaps fairness is achieved by a more distributed or different process (or simply a tendency, not a hardwired mechanism).

We also note: a null in TMS could mean either “the brain region isn’t actually involved” or “our TMS didn’t effectively modulate it.” We’d lean toward the former only after confirming via things like neuronavigation data that we indeed hit the target and engaged it (for example, maybe a Go/No-Go after TMS shows no change, implying rIFG was still fine, meaning our cTBS didn’t do much – then our experiment was inconclusive rather than a true theory test).

12.3.12 Safety, comfort, and reporting

Finally, we commit to transparently reporting all aspects of the TMS intervention:

We will report any adverse events (e.g. if someone had a headache, or in the extremely unlikely case a minor seizure occurred – which we do not expect at all given parameters, but still). Also, we’ll detail each participant’s stimulation intensity (like the % of motor threshold), the coil orientation we used, how accurately we kept on target, etc. This is important for others trying to replicate or meta-analyze.

We’ll share the code and data relevant to the TMS analysis. This includes the exact brain target coordinates, the neuronavigation files if possible, and de-identified trial-by-trial data showing choices under each condition. The idea is to enable others (even skeptics) to reanalyze or verify our findings. As part of open science, our goal is not just to get a cool result but to let others scrutinize it thoroughly.

12.3.13 Where we go next:

The perturbation tests introduced even more complexity to the experimental design. As we proceed, maintaining discipline in design and analysis becomes paramount – otherwise one could cherry-pick positives or overlook confounds. Therefore, before wrapping up the lab investigations, we dedicate Section 12.4 to outlining how we pre-register, power, and standardize these studies, and how we pit our model against alternatives fairly. In other words, how do we ensure that if this idea holds up, it’s not a fluke or artifact, and if it doesn’t, we catch that honestly? The next section is essentially a research checklist and toolkit for doing these lab studies right.

12.4 Research Notes: Preregistration, Power, ROIs

This section is a lab-facing playbook to ensure our experiments are rigorous and reproducible. We cover what to include in preregistrations, how to plan adequate power, exactly where to look in the brain (ROI definitions), and how to handle analysis decisions so that results from different labs can be compared apples-to-apples. The aim is to make our datasets interoperable and our analyses transparent, so that multi-lab efforts can quickly adjudicate whether the Law of Fairness account holds up.

12.4.1 Preregistration essentials (checklist)

A suggested checklist for preregistering these studies:

Hypotheses and contrasts: Clearly state the primary behavioral hypothesis (e.g., “*Short horizon will increase the probability of choosing the high-Φ option, compared to long horizon, at equal immediate payoff*”). Statistically, this is the $H_{cond} \times \Phi$ interaction in a logistic model. Also state the neural/physio omnibus hypothesis (e.g., “*rIFG/ACC signals will show greater activity for low-Φ under short horizon; vmPFC will show a value surplus for high-Φ under short horizon; autonomic measures will show faster settling after high-Φ under short horizon*”). If we plan to test channel moderation, include “*Open vs. Closed channel will interact such that the Φ × Horizon effect is larger when channel is open*”. If perturbation is included, state “*A three-way $H_{cond} \times \Phi \times Stim$ is expected*”. Basically, commit to the key interaction effects as the targets of inference.

Smallest effect size of interest (SESOI): Define what size effect would be considered meaningful. For example: “*Behavior: an odds ratio ≥ 1.40 for high-Φ choices (short vs. long) is of interest*”. Or “*RT difference ≥ 30 ms*”, “*menu entropy drop ≥ 0.20 SD*”. For EEG: perhaps “ *≥ 0.3 Cohen’s d for the difference in rIFG beta power interaction*”. For fMRI: “ *≥ 0.35 Cohen’s d in ROI contrast*”. These numbers might come from prior studies or pilot data. Setting them prevents after-the-fact moving of goalposts.

Nuisance model: Pre-specify the nuisance regressors we will include (immediate utility, risk level, trial number, fatigue rating, etc.). Also list individual-level covariates we’ll account for (age, sex, perhaps baseline anxiety if relevant, time of day tested, etc.). And importantly, note that *the key QS model must retain the $\Phi \times Horizon$ term after including these nuisances*. This ensures we plan to test the robustness of the effect.

Inclusion/exclusion criteria: Define who is in your sample (e.g., adults 18–65, normal or corrected vision, right-handed if we use motor measures). And define data exclusion rules: e.g., “*Exclude participants who have $>30\%$ of trials invalid*” (like if someone didn’t pay attention), or “*Exclude blocks where manipulation check fails (participant did not perceive the horizon difference)*”. Or “*EEG data with $>40\%$ trials lost to artifact will be*

excluded”, “*fMRI runs with >0.5 mm median motion*”, etc. Also state that we will report results both *with all data (intent-to-treat)* and *with exclusions (per protocol)*, to be transparent.

Primary endpoints and decision criteria: Decide in advance what constitutes a success. For example: “*We will deem the experiment successful if the behavioral interaction is significant and at least one of the neural signatures is significant in the predicted direction.*”. Also define what would falsify the hypothesis (e.g., “If the final ledger in a life simulation ends outside a preregistered equivalence band in a well-cared context, that challenges LoF” or in lab terms, “if the horizon manipulation yields no difference in two independent tasks, we will question the effect” – though that’s more for Section 12.6). Having pass/fail criteria avoids bias in interpretation.

Blinding and randomization: Note how conditions will be randomized (block orders, etc.). State that experimenters are blind to condition where possible (especially in TMS studies – the person running the task should ideally not know if the coil is real or sham) and that participants are blind. Plan to assess blinding (e.g., ask participants, “Did you think you got real or placebo stimulation?”).

Data and code sharing: Commit to how data and analysis scripts will be shared. For instance: “*Upon publication (or within 6 months of data collection end), we will upload de-identified trial-level data, task code, analysis code, ROI masks, and the preregistration document to an open repository (such as OSF or OpenNeuro).*”. Also commit to posting the preregistration URL or DOI in the paper for verification.

This checklist ensures anyone reading the preregistration knows exactly what to expect and what each outcome will mean.

12.4.2 Power planning (worked templates)

We provide some template calculations for different parts of the study:

Behavioral interaction: For a within-subject design comparing short vs. long horizon (perhaps averaged over trials) with some correlation between conditions (since the same people do both), we can use standard formulas or simulations. Suppose we expect an odds ratio ~1.4 (meaning 40% higher odds of a compensatory choice when horizon is short). For α and power targets set a priori, required N depends strongly on within-subject correlation, trial counts, and the expected interaction size; in practice, many designs land in the “order of 10^2 participants” range for robust detection of modest odds-ratio shifts, but each lab should simulate its own design rather than reuse a generic N. This sets a ballpark that these aren’t tiny studies.

EEG/MEG interactions (rIFG beta, ACC theta): These signals often have small-to-moderate effect sizes (say $d = 0.30$ to 0.40 for the difference in power between conditions). To detect that, and especially with corrections for multiple comparisons over time or sensors, we might need $N = 50\text{--}70$ people *with a lot of trials each* (e.g. >400 artifact-free trials per condition). Cluster-based permutation tests help, but planning for the higher end of N is wise.

fMRI interactions (vmPFC value, etc.): With ROI-based analysis, if we expect around $d = 0.35$ for the interaction contrast in key regions, then using a standard power analysis for paired t-tests or small ANOVA, we'd need roughly $N = 40\text{--}60$. If resources allow, multi-site pooling can help. Alternatively, one can plan for sequential Bayesian approaches (e.g., collect data until Bayes factor is decisive) – but one must set those rules a priori.

TMS modulation effects: Three-way interactions are even trickier (Horizon \times Phi \times Stimulation). Based on prior literature, TMS often yields small effects. We estimated $N = 36\text{--}50$ for behavior and $N = 24\text{--}32$ if looking at combined TMS+EEG outcomes with strong within-person effects. Essentially, fewer people are needed when each person does many conditions (since each person is their own control), but dropout and variance in TMS response is high, so we pad the N .

Attrition/Exclusion buffer: Always plan +20% or so more recruitment than the bare minimum power calculation suggests. People might quit after session 1, some data might be lost to EEG artifacts, a few might fail the manipulation check, etc. If we need 40 good datasets, maybe recruit 50.

We include formulas or references for these calculations in an appendix, so labs can adapt with their expected effect sizes.

12.4.3 Regions of interest (coordinates, masks, extraction)

To facilitate consistency, we list the ROI definitions for brain analyses:

rIFG ROI: Use an anatomical mask for right pars opercularis + pars triangularis (from a standard atlas like Desikan-Killiany) – essentially Broca's area on the right. Center around MNI [52, 14, 18] with maybe a 6–8 mm radius if using spheres. Additionally, one could specify a couple of alternative peak coordinates from literature (e.g. [50, 20, 6] or [54, 12, 26]) if doing sensitivity checks.

dACC ROI: Use, for instance, the Harvard-Oxford atlas definition of dorsal anterior cingulate (perhaps 25–50% probability threshold) intersected with a 10 mm strip around the midline to focus it. Center roughly at [4, 24, 36] as given. This captures the dorsal/mid-cingulate zone.

vmPFC/OFC ROI: Could use a meta-analytic mask for “value” – e.g., Neurosynth for “reward” or “valuation,” which often yields a cluster in medial OFC. Or just take a sphere around [0, 44, -8] (in frontal medial wall) with some expansion. We mention a specific approach: e.g., take an available parcellation like the Schaefer-400 atlas, select the parcel in vmPFC, and intersect with a meta-analytic map. The key is to define it *before* seeing our data.

Anterior insula ROI: Use Destrieux atlas for the short gyri of the insula and anterior circular sulcus, in both hemispheres. Or simpler, a 8mm sphere at typical coordinates (e.g., right anterior insula ~[34, 22, -2]; left ~[-32, 24, 0]).

Control ROIs: We also define a few “negative control” regions: e.g., primary motor cortex (hand knob area ~[38, -22, 54] on the right), primary visual (like occipital pole around [± 10 , -90, 0]), maybe a cerebellum or left angular gyrus if needed. These are regions we *don’t* expect to show any $\Phi \times$ Horizon effect; including them helps verify specificity (they should remain flat).

For each ROI, clarify how data will be extracted:

For fMRI, we’ll take the mean percent signal change within the ROI for the contrast of interest (short-high Φ vs. short-low Φ vs. long differences, etc.), possibly use robust stats (to minimize outlier influence) and maybe also report medians.

For EEG/MEG source-level, we’ll either project sources into those ROI regions (e.g., use beamforming or equivalent) or at least identify sensor clusters that correspond to them and then confirm with source localization. We may report sensor-level results with topographies, but ROI-based source projections make it easier to compare across studies.

Time windows: We declare in advance which time windows correspond to our primary signals: e.g., 150–300 ms for rIFG beta, 250–600 ms for ACC theta, 300–700 ms for vmPFC value potential, and a post-choice 1–30 s window for autonomic responses. This prevents fishing later – we’re locking in when to look.

By sharing these definitions (perhaps even providing the ROI mask files online), we make it easier for another lab to exactly test the same thing.

12.4.4 Preprocessing pipelines (declare once, reuse)

An often underappreciated source of inconsistency is different data preprocessing. We therefore standardize and preregister those steps:

Behavioral data: We'll remove nonsensical trials (e.g., if someone responded in <250 ms, that's likely a mistake) and extremely long outliers (say >3 SD above mean RT). We'll z-score or otherwise scale Φ values so that model coefficients are interpretable (center them to mean 0, etc.). We will compute things like entropy per block as pre-defined (Shannon entropy of choices across categories, for example).

EEG/MEG: Specify filters (e.g., band-pass 0.1–40 Hz for EEG), how we will handle line noise (notch at 50/60 Hz if needed). Mention artifact removal: e.g. “Perform ICA to remove EOG (eye blink) and ECG components”, or use methods like ASR (Artifact Subspace Reconstruction) to clean big bursts. Define time-frequency analysis parameters: “Use Morlet wavelets for time-frequency decomposition: beta defined as 15–30 Hz, theta as 4–7 Hz”, and how we will detect “bursts” (maybe by thresholding at the 75th percentile amplitude for beta and requiring a minimum duration of X ms). All this is laid out so that if another lab does it, they can replicate the steps.

Also, specify the head model if doing source localization (use a template MRI or individual MRIs, etc.) and how we'll report any forward model assumptions.

fMRI: Declare that we'll use a standardized preprocessing pipeline (maybe mention fMRIprep or SPM with certain steps). E.g., include slice timing correction, motion correction, susceptibility distortion correction if available, and ICA-based denoising (like AROMA) if we choose that. We note that our first-level GLM will include regressors for key events: option onset, choice moment, etc., with modulator variables for Φ , Horizon, their interaction, plus nuisance regressors for motion parameters, physiological noise (via something like aCompCor), maybe reaction time if it varies, etc. We specify the HRF model (say SPM's canonical HRF with a temporal derivative) and high-pass filter (maybe 128 s). And plan for ROI analysis as primary, whole-brain cluster-corrected as secondary (report if any clusters show up outside ROI, but main claims will focus on ROI).

Autonomic (pupil/EDA/HRV): Describe how we will preprocess those signals: e.g., pupil data will have blinks interpolated and a smoothing filter (Savitzky–Golay) applied, baseline subtract using 500 ms pre-choice baseline, then measure peak dilation and a decay constant. EDA will be decomposed (using convex optimization methods or simple peak detection) to get phasic responses per trial and tonic shifts. HRV we'll compute perhaps RMSSD (root mean square of successive differences) in a sliding window, etc. These details ensure consistent measurement.

Having a predeclared pipeline guards against accusations like “you only got that result because you cherry-picked a filtering method or noise removal technique.” If any deviations occur (say we discover a needed change), we document it.

12.4.5 Model set and adversarial comparisons

We must not only fit our QS model but also compare it to plausible rival models to see which explains data better. We outline the set:

QS model (target): This is the model we've been using – includes $\Phi \times \text{Horizon}$, and their interaction (and further interactions with TMS if in that experiment), plus nuisance regressors. Essentially, it assumes compensability and horizon together drive a unique effect.

Risk-only model: A model where decisions are driven by immediate reward and risk (variance or ambiguity) and perhaps effort, but no Φ term at all. If our tasks involve any risk differences, this model will try to attribute choices to risk aversion instead of QS. It would include terms like risk level, maybe squared if needed, and all nuisances.

“Homeostatic” Reinforcement Learning model: This alternative posits that people just try to maintain some setpoint of reward or comfort, without an explicit fairness mechanism. We might formalize it as: value = immediate reward minus some deviation penalty (e.g. if they've had a streak of good outcomes, they're less hungry for more – kind of an adaptation model). Importantly, *no Φ factor here*, just assume people adapt to accumulated outcomes. This model might predict some compensatory behavior as a byproduct of reward dynamics, but not tied to horizon per se.

Pure difficulty/effort model: Perhaps all we are seeing is that near the end people get tired or want to avoid effort, so they narrow choices because of fatigue. This model would include things like time-on-task or trial count, an overall tendency to choose low-effort options increasing as time goes on (which might mimic “menu tightening”). It would predict a uniform focusing effect due to fatigue, not specifically because options differ in Φ .

We can also have variants like a framing-only model where maybe the words “last chance” just cause an emotional reaction but no rational balancing (if that were the case, it might predict some inconsistent patterns, but we include it conceptually).

We then state how we'll judge models: e.g., compare via AIC/BIC, likelihood-ratio tests, or Bayes factors for Bayesian fits. The decision rule being: the QS model should outperform these rivals in fit (lower AIC, higher likelihood, etc.) *and* critically the QS model should have a positive $\Phi \times \text{Horizon}$ term remaining. If a rival fits just as well and explains the data without needing that interaction, then QS hasn't proven itself.

We essentially pre-specify an adversarial collaboration with ourselves: let the alternative explanations fight it out with our model in the analysis.

12.4.6 Multiplicity and stopping rules

Given we test multiple outcomes (behavior, neural, etc.), we plan how to handle multiple comparisons and any sequential analysis:

Family-wise error control: We group outcomes into families: e.g., behavioral family, EEG family, fMRI family, autonomic family. Within each family, we may correct for multiple measures if needed using procedures like Holm–Bonferroni or Benjamini-Hochberg FDR at $q=0.05$. For example, if we have two primary behavioral measures (choice and RT), we control those together. If we have four ROIs in fMRI, we might control across them if they're not fully independent.

Sequential plans: If we opt for Bayesian sequential analysis, we declare upfront a maximum N and stopping criteria (e.g., “*We will collect up to 60 participants and will stop early if at any interim analysis the Bayes factor $BF_{10} \geq 6$ in favor of the interaction, or $BF_{01} \geq 6$ against it*”). Otherwise, we do fixed-N.

Interim looks: If we *do* peek (which is discouraged), it would only be for data quality checks. We pledge not to peek at the group differences early and stop just because it looked significant or not (that inflates Type I/II errors). Any interim checking will be blind to condition labels or only to ensure assumptions are met.

Declaring this prevents ending the study conveniently at a $p=0.049$ or something like that.

12.4.7 Reporting template

We even provide a structured template for reporting results, to ensure nothing is omitted:

Participants: “N = ... recruited; N = ... analyzed (after exclusions). Exclusions: X for failing manipulation check, Y for excessive artifacts, ...”.

Manipulation checks: “Short-horizon condition: mean perceived urgency = ..., long-horizon = ..., $t(df)=...$, $p=...$ (successful). X% of participants correctly identified final vs. not final session, etc.”.

Behavior: “Odds ratio for choosing high-Φ (short vs. long) = ... (95% CI [..., ...]), indicating [increase/decrease]; Reaction time difference = ... ms, etc.”. If using regression coefficients, report those with CIs too.

Neural (EEG/fMRI): “In rIFG ROI, Δ beta power =... ($d = ...$, $p = ...$); ACC theta cluster $p = ...$; vmPFC BOLD difference = ...% signal (CI, p ...)”. We ensure to report effect sizes (like Cohen's d) and confidence intervals, not just p-values, for key contrasts.

Autonomic: “HRV (RMSSD) increased by ... ms (p=...) after high-Φ short-horizon choices vs. others; EDA decay constant difference = ...; pupil dilation: Δ = ...”.

Adversarial models: “Risk-only model AIC = ..., QS model AIC = ... (Δ AIC = ...); likelihood-ratio test p = ... supporting QS’s inclusion of $\Phi \times$ Horizon. Homeostatic model had RMSE = ..., worse than QS by ...”. Basically, we summarize which model fit best and whether adding the QS interaction significantly improved predictive power.

Data/code: “Data and analysis scripts are available in an openly accessible repository with a version identifier; ROI mask files and the preregistration record are provided alongside”.

This template ensures we report *all* critical things consistently (so one study’s report can be lined up with another’s easily). It’s like a checklist embedded in the results section.

12.4.8 Minimal Reproducible Bundle (MRB)

To further encourage open science, we define what to include in a shared repository for full reproducibility:

Think of a folder structure:

/task: All materials needed to recreate the experiment – stimuli images or texts, code for the experiment (if using PsychoPy, E-Prime, etc.), instructions given to participants, and any custom scripts for localizer tasks.

/preproc: The exact preprocessing pipelines as code or containerized environments (e.g., a Docker or Singularity image with all dependencies). If we used EEGLab, provide the EEGLab scripts; if fMRIprep, note the version. This ensures someone else can process the raw data the same way.

/rois: The ROI mask files (in NIfTI for MRI, or sensor lists for EEG) and coordinates. So others know exactly what brain region was analyzed.

/stats: The model formulas, maybe a copy of the preregistration text, the SESOI values used, and multiple comparison procedure details. If Bayesian, list priors. Essentially all the analytical decisions in a form that can be applied to new data.

/data: Anonymized data sufficient to reproduce the analysis. For behavioral, a CSV with one row per trial or per participant-condition. For EEG, perhaps epoched data or at least processed time-frequency data per participant. For fMRI, first-level contrast images (so others can try a meta-analysis or re-run group stats). Because raw fMRI or EEG can be huge, we at least give the distilled form (though raw could be shared via larger repositories too).

/docs: Documentation including a copy of the consent form (if relevant), the preregistration PDF, a flowchart of participant inclusion (CONSORT-style diagram), and a log of any deviations from the plan.

By assembling this “bundle,” any independent researcher could, in principle, re-run our analysis or incorporate our data into an aggregate analysis with minimal hassle.

12.4.9 Cross-site harmonization tips

If multiple labs attempt these experiments, we want to ensure their data can be combined. Some tips (which we also plan to follow):

Use identical horizon framings and instructions wherever possible. If one lab says “this is your last chance” and another says “no more chances after this,” those are similar enough, but we’d prefer literally the same phrasing to avoid any subtle differences. We might even share a short script or video so it’s standardized.

Use the same operational definitions for open/closed channels. E.g., if “repair tickets” are used, make sure every lab defines them and displays them in the same way (number of tickets, how they deplete, etc.).

Exchange pilot data: before fully running, labs can share a small set of pilot results to check if, for example, one lab’s equipment adds a weird delay or if participants interpreted something differently. This can catch issues early.

Possibly do a multi-site mini-meta periodically: say after each lab has 20 participants, pool the data (without looking at outcomes by condition maybe, just overall variance) to ensure no lab is an outlier in terms of data distribution. Or check manipulation check consistency across labs. This can detect if, say, one site’s participants never believed the “last round” story, meaning that site might need to tweak how it’s presented. All without actually peeking at the key effect until the end.

12.4.10 What would count as a “clean null”

We also want to define in advance what pattern of results would convince us that the QS effect truly isn’t there (at least in lab tests), rather than being inconclusive.

A clean null would mean:

Behavioral null: The $\Phi \times$ Horizon interaction in behavior is effectively zero, with the confidence interval falling entirely inside our equivalence bounds (e.g., OR between 0.95 and 1.05) across two independent paradigms that had verified horizon manipulations. In other words, despite doing two good different tasks, we found no hint of a compensatory shift.

No ROI effects: None of the predefined ROI signals show the interaction after proper nuisance regression – and not only non-significant, but with effects so small (say all Cohen's $d < 0.2$) that they're within the “null corridor” we specified.

Rival models fit just as well: The alternative models (risk, effort, etc.) fit the data as well as QS, meaning QS's unique predictions aren't needed to explain anything.

TMS perturbation null: In any perturbation study, rIFG/ACC stimulation failed to alter the horizon effect at all, and even a high-powered multi-lab attempt showed basically nothing (with tight CIs around zero difference).

If *all* of those happened, especially across multiple labs, that would justify a strong skepticism or “non-lawlike tendency” interpretation of LoF. In essence, if careful experiments consistently show nothing where something should be, we have to consider that the Law of Fairness might not hold as a robust law, and perhaps at best people have a mild inclination or it was an artifact of other things.

We mention that a *package* of such nulls from multiple labs would be needed to really declare defeat, because any single experiment could be flawed. But if, say, three labs each did a different task and all got nulls with overlapping zero-effect confidence intervals, then the evidence would lean toward QS not being a thing.

12.4.11 Where we go next:

With the core design and analysis principles set, Section 12.5 moves on to the idea of negative controls in more detail – which we've touched on but now enumerate systematically. These are checks to ensure that wherever the theory predicts *no effect*, we indeed see no effect. They complement what we've been discussing here by adding further ways to catch spurious results. Essentially, after designing for positives, we design for where the effect should not appear. If those places do show an effect, it lights up a warning that perhaps something is off in our experiment or theory.

12.5 Negative Controls

Positive evidence for the Law of Fairness is only compelling if we simultaneously show that the same methods do not produce QS-like effects in situations where the theory says they shouldn't. Negative controls are experiments or analyses designed to yield a null result if our interpretation is correct. They help ensure we're not fooling ourselves with artifacts or confounds. This section lists several categories of negative controls – spanning task design, stimulation sites, signal processing, analysis choices, and interpretation steps – that we include as guard rails. Researchers should implement multiple negative controls in every study and preregister them alongside primary tests.

12.5.1 Task controls: matched difficulty without compensability

These controls use tasks very similar to our main paradigm but remove the compensatory element (Φ) while keeping other aspects similar, to check that horizon by itself doesn't cause some generic change:

(A) Risk-only choice: Design a decision task where options differ in risk/uncertainty and perhaps immediate effort, but not in any future reparative potential. For example, one option might be a safe bet (small guaranteed reward) vs. a risky bet (could win or lose points), with equal expected value. Importantly, neither option affects anyone else or the future – so Φ is effectively zero or equal for both. Expectation: If we impose a “short horizon” frame on this risk task (like “only a few trials left”), we should not see a tilt because there is no compensatory dimension to tilt on. If instead we saw people become more cautious just because of the short horizon, that would suggest maybe the horizon effect we thought was about fairness is actually a general caution effect. This control distinguishes “people become careful near the end” (which is not specific to QS) from “people specifically seek compensatory actions near the end.” We predict essentially no change here, confirming it's not just caution.

(B) Effort-only trade-off: A task where one option requires high effort for a fixed reward and another requires low effort for the same reward – but again, no opportunity to improve future outcomes or others' outcomes in either case. For instance, pedal an exercise bike for 1 minute for \$1 vs. press a button for \$1. In some blocks, we might say “this is your last chance to earn money,” etc., to mimic horizon urgency. Expectation: If horizon per se (with no channel to do good) doesn't create the QS effect, participants *shouldn't suddenly choose the high-effort option more* just because it's the final round – because high effort doesn't repair anything, it's just harder. Any difference would likely be from something else (like maybe they conserve energy at the end?). This control helps ensure our main effect wasn't due to an effort aversion interplay.

(C) Framing-matched decoys: We replicate the main task's exact framing, timing, visuals, etc., but alter the nature of the outcomes such that what was formerly a "repair" option is now just labeled differently but has no real downstream benefit. For example, in a decoy version, the option might still be called "Option A" with a similar description length, but it doesn't actually help anyone (though in the main task Option A did). Expectation: If the QS effect is real, just the *appearance* or label without actual consequence should not produce the vmPFC surplus or choice tilt. In other words, participants won't prefer Option A in the short horizon if it's actually meaningless. If they still do, it suggests maybe wording or placebo could be driving the effect rather than actual compensability.

Overall, task controls rule out the possibility that short horizons simply make people more conservative, more diligent, or any other general effect. They pin it specifically on the presence of compensatory potential.

12.5.2 Site controls: stimulation and ROI specificity

These are especially relevant for the perturbation and neural analysis aspects:

(A) Stimulation site control (e.g. vertex or irrelevant cortex): We include a condition where we apply the same TMS protocol to a brain region not implicated in QS – commonly the vertex (top of the head, roughly over medial parietal cortex) or perhaps a site like the left temporoparietal junction or angular gyrus which we don't expect to play a role here. Participants would do the same horizon task after this control stimulation. Expectation: No change in the behavioral $\Phi \times$ Horizon interaction compared to sham, and no systematic changes in any neural signatures (like beta or theta) since we didn't target those circuits. If we *did* see an effect – say vertex TMS also reduced the horizon effect – that suggests the TMS results were likely due to a general factor (maybe arousal or distraction) rather than specific rIFG/ACC involvement.

(B) ROI control (unrelated brain regions in analysis): In our fMRI/EEG analysis, we always check a couple of control ROIs like primary motor cortex (M1) or primary visual cortex. These regions shouldn't logically show any $\Phi \times$ Horizon interaction if our theory is correct. Expectation: After proper nuisance regression, these control ROIs show flat lines – no special activation differences for high vs. low Φ by horizon. If instead we found a "false positive" pattern in M1 or V1, that might indicate our task conditions correlate with something like attention or difficulty that affects the whole brain, or maybe a failure in how we did baseline corrections. Essentially, control ROIs help confirm that our analysis isn't picking up a widespread effect (like "short horizon makes people generally tense, which increases all brain activity").

These site controls ensure that when we say “rIFG and ACC are special,” we’ve checked that other places didn’t also light up in the same way (which would hint at a more global phenomenon or analysis artifact).

12.5.3 Signal controls: physiological and eye-movement artifacts

Given we rely on physiological measures (pupil, EDA, heart rate) and EEG signals that can be influenced by trivial things, we set up controls to verify that our signals are not mirages:

(A) Blink/microsaccade control: Eye blinks and tiny eye movements can cause apparent changes in EEG (frontal channels especially) and also obviously affect pupil size. We record eye-tracking during tasks, so we can measure blink rate and microsaccade (tiny involuntary eye flicks) rate. Design: We include these ocular metrics as regressors or do an analysis removing segments with blinks to see if effects persist. Expectation: The key neural differences (like ACC theta or rIFG beta) remain after accounting for blink/microsaccade rate, and there is no QS-like pattern in the eye data alone. For instance, it shouldn’t be that people simply blink less when doing a compensatory action under time pressure (which could produce an “increased beta” artifact from eye muscles). And if we just looked at blink counts or so, we shouldn’t see them mimicking our effect (like more blinks for one condition would be a confound). Ensuring QS neural markers aren’t explained by eye movements is crucial, since eyes respond to many things (light, concentration, etc.).

(B) Heart/respiration regressors: Another potential confound is that our conditions (short horizon, stress of compensating, etc.) might differ in heart rate or breathing patterns, which can influence fMRI signals (via blood CO₂ changes etc.) and even EEG to a degree. Design: In fMRI analysis, we use techniques like RETROICOR or include respiratory volume and heart rate variability as regressors. In EEG, we can similarly record ECG and breathing and remove related components. Expectation: After doing so, the QS interactions in neural signals should remain. Also, looking at the raw cardiac/respiration signals alone, we should not see a pattern that mirrors QS (like heart rate jumps only for high-Φ short horizon). There might be global shifts (e.g. everyone’s heart rate might be a bit higher in the tense final block), but it shouldn’t carry the specific signature of our effect. If it did, we might be looking at a general arousal effect rather than QS per se.

Signal controls protect against the critique: “Your EEG difference could just be because people didn’t blink as much when doing X, or your fMRI difference is because people held their breath in condition Y.” We actively show that’s not the case.

12.5.4 Analysis controls: pipelines and model swaps

We want to ensure our findings are not an artifact of a particular data processing or modeling choice. So:

(A) Pipeline swap: We plan at least one alternative preprocessing pipeline for EEG and fMRI and confirm that results hold in both. For example, we might preregister that for EEG we'll try an ICA-first vs. ASR-first pipeline (two ways to remove noise), or for fMRI an ICA-based vs. GLM-based denoising. Expectation: The QS effects (and importantly, the lack of effect in negative controls) should appear in both pipelines. If an effect only shows up in one pipeline and not the other, that's a flag that it could be a data processing artifact. We'd then investigate and report that.

(B) Adversarial model fit (on same data): We already described comparing QS vs. alternative models in 12.4.5. Here it's reiterated as a control: fit simpler models (risk-only, effort-only, etc.) to see if they can produce the omnibus pattern. Expectation: Those rival models should not produce a similar $\Phi \times$ Horizon interaction when we artificially allow them to incorporate horizon or channel after the fact. For instance, if we shuffle the labels of what's "repair" vs. "indulgence" (destroying the true Φ structure), then fit a risk model plus horizon, it should not magically generate the pattern. This is a bit abstract but essentially we check that you can't explain the results by retrofitting another theory's model. If a risk-only model, once you let it also use horizon as a factor, still can't match the QS pattern, that supports QS specificity.

(C) Label permutation: As another robustness check, we might randomly shuffle which options are considered high- Φ vs. low- Φ within the data (preserving other features). Design: For each participant, randomly assign their choices labels as "high" or "low" compensability regardless of actual nature, and then test if any horizon interaction appears (this would be like testing if our analysis would pick up false positives when there is no real alignment). Expectation: On permuted data, the interaction should collapse to ~ 0 . This ensures our analysis isn't somehow biased to find an interaction even in random data (maybe due to imbalance etc.). It's an extra sanity check that the mapping from features to Φ is what drives the effect, not some spurious aspect.

In summary, analysis controls show that our results are not flukes of a single processing stream and that alternative explanations truly fall short even when tested on the same dataset.

12.5.5 Instruction and expectancy controls

These control for psychological factors like demand characteristics or participant interpretations:

(A) Demand-blind framing: We run variants of the experiment with different cover stories to see if participants' knowledge of what we're testing might influence behavior. For example, one group of participants might be told the study is about "time management strategies" and the other group told it's about "risk tolerance," but they actually do the identical task with different framing of purpose. If participants consciously guess we're looking at fairness, they might try to behave accordingly. Expectation: The $\Phi \times$ Horizon effect should not depend on what cover story they hear. If it did (say, only those who thought it was about fairness showed the effect), it could mean the effect was partly due to demand characteristics. We aim for a scenario where even under a misleading or neutral cover story, the effect emerges, indicating it's not just participants trying to please the experimenter.

(B) Belief-about-stimulation covariate: For TMS studies, as mentioned, we ask participants after each session whether they think they got real or sham stimulation. Design: Include this belief as a covariate in analysis (or even look at a 3-way interaction with that). Expectation: The TMS effect (difference between active and sham) should remain after accounting for whether participants thought it was active. Also, in the vertex control sessions, even if participants thought it was active TMS, it should yield no effect. If someone's mere belief they were stimulated could produce a change, then we have a placebo effect to consider. Ideally, we see that *even those who thought they were in sham but were actually stimulated show the effect*, proving it's physiological, not just expectation.

These expectancy controls make sure what we see isn't just people doing what they think we want. The Law of Fairness posits a deep-seated mechanism, not a conscious strategy to appear fair.

12.5.6 Channel manipulation controls

We double-check that it's truly the reality of compensation driving things, not just the *instruction* about it:

(A) "Fake open" channels: We announce that a compensatory route exists, but in the task logic we secretly block it. For instance, tell participants, "If you press this, you'll send relief to your partner," but under the hood that action doesn't actually do anything beneficial (unknown to them). Or maybe more ethically, we can simulate something where they *think* it's helping but we analyze separately the ones that would vs. wouldn't. Expectation: If participants merely *believe* a channel is open but it's not actually effective, we might still see some behavior change, but we predict it will be diminished or absent when outcomes reveal no benefit. Essentially, if it's only the expectation that

matters, they might still shift behavior; but if actual efficacy matters, then a fake channel shouldn't amplify QS signatures. We suspect actual affordance reality is key (the system might subconsciously detect if the actions are futile, e.g., no reward prediction error improvement, etc.). If fake open does nothing, that confirms it's the true outcome that counts.

(B) "Fake closed" channels: The reverse: we tell them no compensatory route is available, but in fact there is one embedded. E.g., instruct "no matter what you do, you can't help your partner in this block," but actually if they choose a certain option it does help. Participants might not initially try since they think it's closed, but if some do or they sense the effect, we observe behavior. Expectation: Even under this pessimistic instruction, *if* the affordance is truly there, over trials they might realize or feel that their actions have impact (maybe they see partner outcomes improving) and thus the QS effect should still emerge, though possibly delayed. This checks whether it's purely instruction-driven. Why do this? To show that it's the actual existence of a route (which participants can sometimes feel via subtle feedback) that drives the effect, not just what we tell them. If QS is a fundamental mechanism, it might operate even if the person was initially told nothing can be done, as long as in reality something can.

These are a bit tricky ethically (we have to deceive participants), so they'd be done carefully and debriefed. But they really stress-test whether QS is about actual outcome contingencies or just beliefs.

12.5.7 Temporal controls: horizon without endgame

We ensure that it's not just any temporal cue, but specifically the combination with compensability, that matters:

(A) Distant-deadline cue: Present a salient countdown or "time is passing" signal that does not actually restrict compensability. For example, a timer is on screen but the participants know it's unrelated to their opportunities (like "this is for pacing only, it will reset"). This can test if a countdown by itself triggers a stress or focus response. Expectation: Simply seeing a timer or having a deadline unrelated to the task's compensability should not produce the QS behavior pattern. rIFG/ACC shouldn't spike just because a timer is there – unless that timer signals loss of opportunity. If we find people get twitchy just due to any countdown (a possibility: some folks just hate timers), then we must ensure in our main tasks that the effect is beyond that baseline.

(B) Optional-stopping surrogate: Instead of a fixed horizon, let participants end the task early if they want, with equal payoff either way. This is like a fatigue test: if someone is tired or bored, they might quit early. We compare short vs. long horizon conditions in a

scenario where they can stop at any point (so horizon isn't externally imposed). Expectation: Without an endgame pressure, stopping rates should reflect generic fatigue, not a QS pattern. If horizon effect was confounded with fatigue in our main study, this control might show something; but under neutral conditions, we expect no targeted tilt. This control is a bit abstract, but essentially if in short horizon blocks people stop sooner (like "oh well, end it now") but in long horizon they continue, that's a different effect – likely not QS but boredom.

12.5.8 Pharmacological and lesion controls (where ethical/available)

These are more rare, but if possible:

(A) Non-specific arousal: e.g., caffeine vs. placebo: Give participants a dose of caffeine (which raises arousal) and see if that mimics QS or not. Expectation: Heightened arousal globally might speed responses or increase heart rate, but should not selectively create a $\Phi \times$ Horizon interaction. If it did, then maybe our effect was just arousal. We expect caffeine might reduce overall RT, but not change the pattern of preferring high- Φ under short horizon.

(B) Focal brain lesions: If data from neurological patients are accessible (e.g., someone with damage to rIFG or ACC), see if they exhibit a weaker QS pattern. Real lesions are like nature's TMS. Expectation (if QS is real): Patients with rIFG or dACC damage might *not* show the usual compensatory behavior as strongly (they might treat last rounds more normally), whereas patients with lesions in unrelated areas (occipital, etc.) would still show it. This is a bit beyond typical lab work, but we note it as an ultimate test: if QS is fundamental, damaging its hardware should dampen it. We also caution that ethical and practical issues limit these tests, but mention them conceptually.

12.5.9 Measurement-only sessions

We consider if measuring without action yields anything:

(A) Passive viewing: Have participants *view* the same options and scenarios but not actually make choices, just observe. Perhaps they predict what they'd do or just watch. Expectation: Without the need to choose, the strong QS neural signals (like rIFG braking, etc.) should attenuate or vanish. If we still saw differences in, say, vmPFC or ACC just from viewing, it suggests some confound (maybe images differ systematically). This also separates decision-bound activity from mere stimulus-bound activity. Ideally, insula and ACC signals are much weaker when they're not actively making choices – confirming those signals we focus on are indeed tied to decision-making under horizon pressure, not just looking at images of needy people or something.

12.5.10 Predeclared null zones (SESOI-based)

We formalize the concept of “null result”:

For each family of outcomes, we set a null zone (based on SESOI). For example, we might say any effect size $|d| < 0.20$ or odds ratio between 0.97–1.03 is effectively zero. We preregister these as thresholds for considering an outcome null.

Decision: If a negative control shows an effect outside this zone (meaning something happened when nothing should have), we *flag it*. For instance, if our risk-only task unexpectedly shows an odds ratio of 1.5 for short horizon (which it shouldn’t), that means our experiment’s specification might be off – perhaps participants misunderstood something or we inadvertently gave that task a compensatory angle.

The procedure then is: if a negative control “lights up” beyond allowed bounds, we stop and recalibrate before interpreting any positive results. It’s like a circuit breaker: do not trust the main findings until you explain/fix why a control failed.

12.5.11 Cross-laboratory portability controls

If multiple labs are involved, we include controls to ensure results generalize across them:

(A) Instruction tape swap: As a neat idea, Lab A uses Lab B’s recorded instruction video, and vice versa. This ensures differences aren’t due to how one experimenter talks or slight wording differences. Expectation: QS positive effects travel, and negative controls remain null, regardless of who’s instruction set is used. If a particular phrasing is crucial, that’s not robust – we want the phenomenon to be robust to minor surface differences.

(B) Stimulus set swap: If labs used different stimuli (images, scenarios), swap them to see if results hold. For instance, if one lab’s task used social scenarios and another used health scenarios, each tries the other’s. Expectation: Effects persist insofar as QS should be domain-general, only tied to compensability, not the exact content (assuming compensability is present in both sets). Negative controls also remain null in both sets.

12.5.12 Synthetic-agent controls (simulation falsifiers)

We even test our analysis pipeline on fake data generated by alternative models:

Design: Create simulated agents that follow strategies like risk aversion, or pure homeostatic adaptation, or always random, etc., with no QS mechanism in them. Feed their “behavior” through our analysis pipeline (as if they were participants).

Expectation: These synthetic agents should not produce the full triad of QS signatures. For example, a risk-averse agent might avoid high-variance options more when horizon is short if they get nervous, but that wouldn't necessarily create the rIFG/ACC vs. vmPFC pattern we're looking for. We anticipate these faux datasets fail to show the combination: e.g., maybe they show menu narrowing but no insula settling, or vice versa, but not all.

If our pipeline somehow “finds” a QS-like pattern in data known to have none (e.g. random choice agents), then our analysis might be overfitting or artifact-generating – that’s a big warning sign. It’s essentially a dry run: the pipeline should only light up when real QS behavior is present.

12.5.13 What to report when a negative control “lights up”

Despite best efforts, sometimes a negative control will show an effect (false positive). We outline what to do in that case:

Immediate actions: If a negative control yields a non-null result:

Tighten nuisances: Add more covariates or remove possible confounds that might have caused it (e.g., maybe our risk-only task still had slight effort differences that accidentally gave a pattern – we then include an effort regressor or redesign it).

Check data quality: Make sure it’s not an artifact (like all high-Φ trials happened to be at the end and participants got tired, etc.). Look at motion, blinks, TMS coil issues in those runs.

Inspect compliance: See if participants misunderstood a condition. E.g., maybe in the “no compensability” block, some still *thought* they could compensate (so from their perspective it wasn’t a true negative control). Use manipulation check responses to verify that.

Outcomes (interpretation):

If after doing the above, the signal/effect disappears, we conclude it was likely an artifact or something fixable. We document it, but proceed with guarded confidence.

If it persists across attempts and even appears in multi-site repeats, then it’s potentially theory-threatening. In that case, we may need to broaden our rival explanations or refine the theory because something is showing up where it shouldn’t. For instance, if every time we run a risk-only task, we still see a little compensatory behavior, maybe people inherently see any decision as an opportunity to do something meaningful (or maybe our definition of compensability needs refinement).

Essentially, a persistent failure of a negative control could mean our idea of QS might be partially wrong or incomplete – perhaps fairness is entangled with other factors we assumed independent. In such a scenario, humility dictates updating the theory or method.

12.5.14 Where we go next:

Now that we have established how to carefully run experiments and what must *not* happen, we turn to summarizing Fail patterns – concrete data outcomes that would count strongly against the Law of Fairness. Some we've already implied (like no horizon effect at all, or a rival model explaining everything). Section 12.6 compiles these into a checklist of “kill switches” for the theory. It's effectively the other side of the coin from our predictions: if these patterns show up reliably, a fair-minded scientist should consider the LoF hypothesis falsified (or at least in need of major revision).

12.6 Fail Patterns in Lab Tests

The Law of Fairness (LoF) and its proposed Queue System (QS) mechanism make specific, risky predictions. This section catalogs the data configurations that, if observed reliably and under rigorous conditions, should count *against* the theory. These are essentially the failure modes that would force us to doubt or abandon the LoF in its strong form. We emphasize treating these as kill switches, not mere anomalies – i.e. if they happen and can't be explained by experimental error, the theory doesn't get to wiggle out easily. We also include notes on how to double-check each pattern (to rule out artifact before concluding the theory is at fault).

12.6.1 Core behavioral nulls (first-order failures)

These would strike at the heart of the behavioral claim:

No horizon interaction: The most direct failure would be if our experiments found essentially no difference in behavior between short-horizon and long-horizon conditions. *Signature*: The preregistered $\Phi \times H_{\text{cond}}$ term (compensability by horizon interaction) is effectively zero – for instance, the odds of choosing a high- Φ option is the same whether the participant thinks it's the last round or one of many. We'd want this observed in more than one paradigm with solid manipulation checks to take it seriously (maybe one could be fluke, but two well-done studies with 0 effect is a big red flag). *Why it matters*: QS requires that horizon length influences the admissible set; if shortening the horizon doesn't actually change choices, then the core idea of LoF driving behavior is undermined.

Triage if seen: Ensure participants truly noticed the horizon difference (if they didn't, the test was invalid). Ensure there actually were opportunities to compensate (if not, then of course nothing would happen). But if all that was in place and still no effect, LoF's behavioral basis would be in question.

Equal tilt when channels are closed: Suppose we *do* see people change behavior near the end, but they do it even when no compensatory actions are available. *Signature*: In a “closed-channel” block (where we removed any meaningful difference between options), horizon still produces some systematic bias (maybe people become generally more conservative or random, etc.). If the tilt towards certain options under short horizon is just as strong even when those options carry no extra future benefit, that suggests the effect is not about compensability at all – maybe it's something like end-of-task anxiety or wanting to finish on a high note, etc., unrelated to QS logic. *Why*: QS specifically posits the mechanism kicks in to balance outcomes; if it fires in contexts where there's nothing

to balance, then our interpretation is wrong (maybe it's not a fairness mechanism, just a quirk of decision-making).

Status: We'd view this as theory-threatening if replicated with clean controls. It would mean we need to attribute the horizon effect to a more generic psychological phenomenon (like a "closing time" effect) rather than a fairness-driven one.

No preference for high- Φ when Φ is matched (symmetric tilt): Another scenario: imagine we set up choices where both options have equal Φ (both are equally reparative or indulgent), just different content. If near the end people don't particularly favor the "repair" one (since they're equal in repair), that's fine. But a failure pattern would be if in a scenario where one option has higher Φ , people *don't* favor it under short horizon. Even worse, if they showed a symmetric or opposite tilt (e.g., they end up just as often choosing indulgence or even more indulgence under short horizon), that directly contradicts LoF. *Signature:* In a head-to-head of a clearly high- Φ vs. low- Φ choice (with equal immediate payoff), participants at end-of-horizon do not lean more to high- Φ , or lean the wrong way. *Why:* That's basically the compensability preference failing to materialize. It would mean the "feasibility of compensation" construct didn't drive decisions as hypothesized.

Note: We must ensure our Φ measure is valid (maybe we mis-estimated which option is truly more reparative). But if we did, say, participants themselves rated one option as more meaningful yet still didn't choose it more at the end, LoF would be in jeopardy for that context.

12.6.2 Neural nulls (hub-level failures)

Even if behavior shows something, the neural predictions could fail:

No hub selectivity after nuisance modeling: We expect rIFG and dACC to selectively activate for low- Φ under short horizon. A null would be if, after removing generic factors, rIFG/dACC show no special interaction. *Signature:* In analysis, once we control for things like difficulty, utility, etc., the difference in rIFG/ACC activity between, say, low- Φ short vs. others is tiny (effect size < .2). That means those regions aren't doing the targeted braking as we thought. If replicable, it weakens QS's neural story strongly.

vmPFC value surplus collapses or reverses: If vmPFC doesn't show higher activation for compensatory choices at equal reward (or if bizarrely it's higher for indulgent choices under short horizon), that undermines the idea of a "fairness bonus" being encoded. *Signature:* No difference in vmPFC BOLD or value signal for high- Φ vs. low- Φ when controlling immediate payoffs, or worse, a negative difference. *Why:* That would mean the brain's value computation isn't aligning with LoF's supposed pressure – maybe

people choose differently for some other reason, not because their valuation system changed.

Distributed, non-specific activation: If instead of a focused network, we see a *diffuse whole-brain response* with no specificity (lots of areas light up but not in a pattern, and ROI differences vanish when controlling motion, etc.), that could imply our measurements are confounded by general arousal or movement. *Signature*: The contrast of short vs. long horizon shows widespread activation in fMRI, but when you check specific ROIs like rIFG, they aren't distinct from global effects, and once you regress out things like heart rate or motion, nothing QS-specific remains. *Why*: That suggests no targeted QS mechanism, just a broad "state change" like being generally stressed as the task ends. That guts the claim of a specific fairness mechanism.

12.6.3 Perturbation nulls (causal failures)

The causal tests could fail in two major ways:

TMS does not modulate pruning: If we dial down rIFG/ACC and nothing happens – participants behave identically to sham – that's a blow to the hypothesis those hubs are critical. *Signature*: No significant difference in the $\Phi \times H$ interaction between sham vs. active stimulation (confidence interval crosses zero effect). We even powered well and saw nothing. *Why*: It weakens the claim that those brain regions implement the QS control. Possibly QS is elsewhere or not causal.

Control sites have same effect as target sites: If stimulating a non-QS region (vertex/IPS) changes behavior *just as much* as rIFG/ACC stimulation does, then any changes we saw might have been due to non-specific factors (e.g., the noise/tactile sensation of TMS, or expectations). *Signature*: The three-way interaction (Horizon \times Φ \times Stimulation) is similar in magnitude for a control site as for the target sites. *Why*: This points to our TMS effects not being specific; maybe people acted differently just because "their brain was zapped" or they thought something happened, not because we targeted QS circuitry. That undermines the evidence for QS being the driving cause.

12.6.4 Autonomic nulls (settling failures)

These target the prediction about emotional "relief" after compensatory choices:

No settling advantage for high- Φ choices under short horizon: If our data show that after a compensatory action, people's physiological arousal (HRV, EDA, pupil) is *no better* or not faster in recovery than after a trivial action, even when time is short, then the idea that the system "feels relief" when doing the right thing near the end is not supported. *Signature*: HRV changes, EDA decay rates, etc., are equivalent for high- Φ vs. low- Φ

decisions at horizon, once you control for effort and duration. Or if high- Φ doesn't lead to any more calmness than low- Φ .

This would suggest that the body's interoceptive reward for balancing isn't there, implying maybe LoF doesn't have a built-in "relief" signal as posited. (It could still be purely behavioral without visceral feedback, but many theories lean on that feedback.)

12.6.5 Model-comparison failures (explanatory displacement)

These would occur if an alternative model fully explains the results:

Rival models match or beat QS on predictive performance: If, for all our measured families (behavior, neural, autonomic), a simpler theory (risk aversion, adaptation, etc.) can fit the data just as well or better and the QS-specific terms add nothing, then LoF loses its necessity. *Signature*: e.g., AIC of a risk+fatigue model is \leq AIC of QS model across behaviors and neural outputs; adding Φ and horizon terms doesn't improve log-likelihood significantly.

Why: If the data can be compressed/predicted without invoking a fairness constraint, then we can't claim a law-like principle is needed. At best, LoF might be just an emergent side-effect of other processes (which means it's not a fundamental law, just a byproduct).

Permutation resilience: If we randomly assign " Φ " labels or otherwise scramble what should be the key variable and our analysis still finds interactions (or a comparable level of fit), it means our pipeline might be overfitting noise or picking up on something else. *Signature*: Shuffling which trials are "high- Φ " yields similar model performance or patterns as the real labeling. That's a very damning sign, as it suggests the supposed structure (Φ) isn't actually what's driving the patterns – maybe something like trial position or visual differences are.

If either of these occur, the case for LoF being a real scientific law weakens drastically.

12.6.6 Robustness failures (pipeline dependence)

We consider it a problem if results only appear under one specific analytic method:

Pipeline fragility: *Signature*: Our core findings disappear or change sign if we alter reasonable preprocessing or analysis choices. For instance, only one particular EEG filtering method produced the effect; using another equally valid method did not. Or the fMRI result only appears with one type of noise regression. This suggests the effect might be an artifact of that method. *Why*: A true phenomenon should be detectable under multiple reasonable approaches (maybe with some sensitivity differences, but not

fundamentally there vs. not-there). If it's that fragile, one worries it's not real or is too context-dependent to call a law.

Instruction dependence without behavior change: If simply telling a different story (like our demand characteristic control) flips the neural or physiological effects while behavior stays same, it suggests those “effects” were demand-driven. *Signature*: Under Cover Story A we see big ACC activation, under Cover Story B we don’t, yet both had similar behavior distributions. That implies the brain differences were not due to QS but due to how people framed the task mentally. That undermines a mechanistic interpretation; it becomes a narrative or semantic effect.

12.6.7 Cross-paradigm and replication failures

These look at generality:

No travel across tasks: If one paradigm yields positive results but a conceptually similar one does not, QS might be overfitted to a particular scenario. *Signature*: e.g., we get the effect in a charity-donation game but not in a helping-teammate game, even though both were set up to test the same principle. If this repeats (each lab can only get it in their special task), maybe what we thought was a general law is actually task-specific. *Why*: A true law should generalize across implementations; failure to travel suggests maybe subtle confounds in the “successful” demonstration or that it doesn’t generalize as claimed.

Multi-site null with harmonized materials: The gold standard hit: if a coordinated multi-lab replication attempt (with good power, same protocol across sites) finds null results (and tight confidence intervals showing any effect is smaller than SESOI), that would be a major blow. *Signature*: A meta-analysis of those labs yields an effect size near zero with CI within our null region. *Why*: It’s hard to argue with that – it suggests either the original was a false positive or context-dependent. For a law-level claim, failing at multi-site is near-fatal.

12.6.8 Dream and sleep adjunct failures (bridge checks)

No valence inversion after tough days (in dreams): If one extension of LoF is that after very negative days people have more positive REM dreams (a proposed low-cost counterweight), but studies find no such pattern, that weakens the broader theory. *Signature*: After a day with low HCl (bad mood/pain), dream reports are not more positive than baseline or don’t reduce next-day distress. If carefully measured and null, it suggests one mechanism (dreams balancing mood) doesn’t hold, challenging QS’s role in sleep.

This is more of a bridge to long-term studies (Chapter 13), but included as a “adjunct Fail pattern.” It’s somewhat tangential to lab tests, but we mention it for completeness of things that could refute LoF in other domains.

12.6.9 Decision thresholds for downgrading the claim

We specify criteria for when we’d “downgrade” LoF from a law to maybe just an observed tendency, or reject it outright:

If any two of the following major failures hold across two well-done paradigms, we’d say the evidence is leaning that LoF is not a guaranteed law but maybe at best a common tendency: (1) behavioral interaction ~ 0 (no effect), (2) no hub-level neural interactions, (3) rivals match QS model. Two strikes, and we move from “LoF is a law” to “maybe people often do this but not reliably or mechanistically.”

If, in addition, TMS fails to modulate and negative controls are all fine (meaning our experiments were valid but still nothing happened), then we’d reject the QS mechanism entirely. LoF might survive only as a philosophical idea (maybe people do tend to even out in life but via other means).

A comprehensive rejection would come if a big multi-site study as mentioned returns null results with high precision, and even the “dream” adjunct (the long-horizon evidence) fails. At that point, we conclude the strong form of LoF (as a constraint on experience) is not supported by empirical data.

In plain language: if careful experiments repeatedly show nothing where something critical was predicted, we either demote LoF to a much weaker notion or drop it. We won’t keep adding epicycles to save it if key tests refute it.

12.6.10 What to do before calling a failure

However, before pulling the plug, we must ensure the failure is real and not due to some oversight:

We outline a checklist to tick off whenever a result suggests a Fail pattern:

Confirm manipulation: Double-check that participants truly experienced the intended conditions – e.g., horizon perception, channel availability, comprehension. A failure isn’t a fair failure if, say, half the participants didn’t realize it was their last round (so there was no actual horizon manipulation).

Recompute Φ : Ensure our calculation of compensability (Φ) was reasonable. Maybe we scored an option as high- Φ but participants didn’t see it that way. We should run sensitivity analyses: if we tweak how Φ is computed or weight things differently, does an

effect appear? If an effect appears under one plausible Φ definition but not another, then maybe our original Φ measure was flawed. We should consider that before abandoning the concept.

Tighten nuisance controls: If a fail could be due to some unaccounted factor (arousal, drift in mood, expectancy), we try adding those into the model to see if anything was masking the effect. Maybe after accounting for something, a small effect emerges (then it's not a total fail, just weaker than thought).

Run negative controls (again): Ensure that in the data, things that should be null are null. If even negative controls misbehaved, the whole experiment might be suspect.

Independent reanalysis: Share the data publicly and maybe invite an “adversarial collaboration” team to analyze it. Maybe they find an effect our analysis missed or confirm the null. This helps rule out analysis bias or errors.

Only after these steps, if the pattern still looks like a failure, do we accept it as such.

12.6.11 Where we go next:

Up to now, our tests have been in controlled bursts of time (minutes, hours, days). But the Law of Fairness ultimately is about lifetimes and long arcs of experience. Thus, Chapter 13 shifts to telemetry-scale studies – tracking people over months or years to see if the predicted balancing acts and compressions occur in the wild. We move from the lab bench to the “long view,” knowing that if LoF is true, it must manifest in real life’s ups and downs, not just lab tasks. The final part of our journey lays out how to observe or refute LoF in longitudinal data, which is the ultimate proving ground for a principle about life’s ledger.

Chapter 13 — The Long View: Telemetry Across Years

Laboratories are great for studying hours or days; the Law of Fairness lives or dies on the scale of years. If the Queue System (QS) truly shapes the menu of thinkable actions to keep lifetime ledgers near neutral, for the vast majority of people, then slow drifts, seasonal cycles, major life transitions, and the long shadow of mortality must all leave footprints that a short experiment would miss. This chapter lays out a practical, ethical, and statistically rigorous program for multi-year “telemetry” – continuous or repeated measurements – to test LoF where it matters most: in real lives, over long horizons, under real-world conditions.

What do we mean by telemetry here? Think of it as fitting life with a gentle array of sensors and check-ins that quietly record how someone’s experience evolves over months and years. With modern technology, this is now feasible. Smartphones and wearables can gather aspects of our behavior and physiology 24/7, and occasional surveys or interviews can capture internal states – all in a way that participants can accept and even find beneficial. The idea is to observe patterns like those we’ve theorized (horizon effects, ledger balancing moves, etc.) unfolding naturally across time, not just in special circumstances.

We focus on three pillars in this long-view approach:

- Measurement: Build a longitudinal *Hedonic Composite Index (HCI)* that aggregates multiple channels of data about a person’s well-being over time. No single measure (not steps, not heart rate, not self-reported mood) is perfect or sufficient, so we integrate many signals – self-reports, physiological data, behavioral logs, even periodic clinical assessments – into one composite metric. Crucially, we do this in a way that no one channel dominates or biases the index. (For example, if a wearable fails for a week, the HCI can still rely on mood reports; if someone under-reports their feelings due to stoicism, maybe their sleep and activity patterns fill in the picture.) It’s like having multiple thermometers in different rooms – we fuse them to get a stable reading of the overall “temperature” of experience.
- Modeling: Use advanced time-series analysis and statistical modeling to detect the patterns LoF predicts: ledger drifts, horizon effects, menu tilts, etc., while separating those from confounds like adaptation or external events. For instance, we might use state-space models or hierarchical Bayesian models to track a person’s latent “affective state” over time and see how it responds when certain thresholds are hit (like a big drop in L(T) or a shrinking horizon due to age or illness).

The modeling also involves creating null models – e.g. standard adaptation theory models – and seeing if those can explain the data, to ensure that if we claim a fairness effect, it's truly because other explanations failed.

- Governance: Implement strong ethical and privacy practices so that this multi-year data collection is respectful, secure, and beneficial to participants. We treat *consent as an ongoing process*, not a one-time form. We design the data handling pipeline to be privacy-by-design (encrypt data, analyze mostly aggregated or on-device, etc.). And we make sure participants get something out of it – whether it's feedback about their own trends, tools that help them reflect, or resources that help them in hard times. The study should be something people are glad to be a part of, not just a monitoring scheme.

Why bother with such a long-term study? There are three key reasons:

First, ledger dynamics are auto-correlated and long-range. A bad month can cast a shadow on the next month's choices; a period of elation can make one complacent until reality checks in. These feedback loops and delayed effects can only be observed if we track continuously. If LoF is at work, it's not just day-to-day fluctuations; it might be that, say, after 6 bad months, something kicks in that wouldn't after just 1 bad day. We need to catch those cumulative effects.

Second, horizons expand and contract naturally over a lifespan. The same person goes from being a teenager (with seemingly endless time) to perhaps a new parent (suddenly life's priorities shift) to maybe a patient with a serious diagnosis (time feels limited). By tracking individuals through these phases, we can see if *the same person* starts exhibiting more “fairness-seeking” behavior as their subjective horizon shrinks. This within-person change is powerful evidence because it controls for personality – you're your own control.

Third, end-of-life isn't an instant – it's a chapter. The process of nearing life's end often unfolds over months or years (think of chronic illness or very old age). This gives a window where, according to LoF, the balancing should accelerate or intensify. Capturing that requires following people into (and ideally through) that phase, in a way that doesn't intrude but observes. By the time someone is in their final week, it's too late to start a study – we need to have their baseline from before and watch the approach. Longitudinal telemetry can provide that continuous storyline.

So, what counts as evidence for LoF at a multi-year scale? It won't be one single “Eureka!” graph. It will be a family of patterns that consistently show up across different people and contexts. Concretely, LoF predicts a few signature phenomena:

- Ledger-drift compensation: When a person's cumulative ledger $L(T)$ goes strongly negative (meaning they've had an extended run of suffering or stress), QS predicts that relief and repair options will become more available and appealing. In life terms, that might mean after a prolonged rough period, we see the person spontaneously gravitate to more rest, seek help, or make life changes aimed at relief – beyond what you'd expect from just getting tired or external rescue. Conversely, after a long positive/excess run (say things have been unnaturally good or easy for a while), we might see the person subconsciously pump the brakes – perhaps taking fewer risks or abstaining from indulgence as if avoiding overshooting. This is a ledger-drift signature: extreme deviations invite self-correction moves.
- Horizon effects in life decisions: As a person's perceived remaining life horizon H_t diminishes – due to aging, a health scare, or even a life milestone that makes the future feel shorter – we expect to see their “menu” of activities and goals narrow and tilt towards those with high Φ (compensatory value). For example, in longitudinal data we might catch that when someone reaches a certain age or stage, they start prioritizing family connections or completing unfinished business more than before, even if externally nothing else changed. Or if someone recovers from a near-death experience (horizon suddenly went to zero and back), afterward we might see a spike in closure-oriented behavior. This horizon signature would appear as measurable shifts in behavior patterns tied to changes in perceived time left.
- End-of-life hedonic compression: As discussed, near the end of life, LoF predicts a compression of affective variance – emotional highs and lows soften and converge toward neutral. Long-term data would show that an individual's daily mood variability (standard deviation) gets smaller as they enter their final phase, compared to earlier in life. Some may have “spikes” of intense positive or negative experiences near the end (like a final meaningful event), but overall, the swings dampen. The net trend of their HCI should level out (approach zero slope) as conscious life closes. We'd look for this pattern across many individuals, controlling for heavy medication effects. We call this end-of-life compression.

All these patterns must stand up to rigorous scrutiny. Life data is messy, so we will account for a host of confounding factors: changes in income, seasons and weather, major world events (pandemics, anyone?), personality traits, known psychological interventions, etc. Only if these signatures remain after controlling for such factors can we attribute them to LoF and not to simpler explanations.

What kind of signals will we collect to detect these patterns? We propose a rich but feasible multi-modal dataset:

- Self-reports (high frequency, low burden): Participants will provide micro-surveys perhaps 1–3 times a day. Each survey is very brief: for example, they could slide a finger on a mood scale (from very bad to very good), rate their energy or stress, note pain or discomfort level, and answer a question like “How far ahead does your life feel right now?” (with options like “just getting through today” versus “I’m planning years ahead”). We’ll also have them check which key “channels” for well-being are available at that moment – e.g. “I have access to: [comfort (pain relief, etc.), companionship, quiet rest, a sense of purpose]”. These daily self-reports act like the ground truth of how they feel and perceive their horizon. Additionally, maybe once a week, they write a short journal entry or voice note – unstructured, just to give context or any narrative they want to share. These are optional but often valuable for interpreting the quantitative data.
- Behavioral patterns via passive sensing: With consent, we use smartphone and wearable data to infer aspects of daily life. Key examples: *Physical activity* (step counts, general mobility range per day – did they stay in one place or move around widely?), *Sleep-wake cycles* (when and how long they sleep, from accelerometer and perhaps sound; disturbances at night), *Communication* (how often they interact with others via calls or messages – we collect metadata, not content, focusing on frequency and breadth of social contact), *Schedule and productivity* (if they use digital calendars or to-do lists, we can see how many tasks they complete or postpone, with privacy safeguards). From these, we derive proxies like “menu entropy” – the variety of different activities or contexts a person engages in per day. High entropy might mean a very varied day (work, gym, social, hobby all in one); low entropy might mean a very monotonous day or a singular focus. LoF would predict lower entropy when someone’s in a compensatory mode (focusing on a few important things).
- Physiology and health metrics: Wearables can also give us continuous heart rate and heart rate variability (HRV), which relate to stress and relaxation; possibly skin conductance for arousal if devices support it; body temperature trends; and so on. We also gather any readily available health data: weight changes, blood pressure (some people track these), medication adherence (if they log it), etc. Sleep quality metrics (like how much deep vs. REM sleep as estimated by a device) are recorded. Additionally, in one week per month, we might ask participants to log any remembered dreams each morning (via a quick voice note)

- this ties into Chapter 10’s theme, looking for “counterweight” dreams across life events. These physiological and sleep measures help indicate how the body is responding; for example, HRV is often higher during restful states, though it is influenced by many factors and is not a direct readout of well-being on its own. We might see patterns like HRV changing following periods where people make compensatory moves, etc.
- Sparse neural “anchors”: We don’t expect people to wear EEGs daily, but as an optional add-on, we might have participants use a simple EEG headset at home once every few months. Or a subset of willing participants might come into a lab or imaging center annually for a short fMRI session. The purpose is to tie our “field” measures back to known neural markers. For example, a quarterly home EEG could involve a 10-minute task that gives us a P300 amplitude (brain’s response to novelty) or an error-monitoring signal, which we might correlate with how their behavior changes in life. An annual fMRI could test whether key regions (like rIFG, ACC) activate strongly when making hypothetical choices under time constraints, and whether that correlates with their real-world pattern of choices. These neural snapshots are not to hunt new effects, but to ensure that, say, a person who shows strong LoF-like behavior also shows the expected brain signatures (if feasible). We treat these neural measures as correlational calibrations, not as proof that any circuit “enforces” QS. They serve as calibration points for our composite index, anchoring the abstract HCl changes to something biologically concrete.
- Contextual life events log: We ask participants (with full control over what they share) to report major life events. We also try to verify these through public data or medical records if possible (with consent). Events include things like: new diagnosis, hospitalization, death of a loved one, marriage/divorce, retirement, job loss, moving to a new city, economic windfall or crisis, etc. These events are important “shock points” where the ledger or horizon might change abruptly. We will analyze data around these events to see how the system reacts. For instance, if someone’s spouse dies (major negative shock), do we see a compensatory pattern kick in over subsequent months? Or if someone retires (horizon suddenly shifts from work-life to end-of-life perspective), do we see them reorient goals accordingly?

We plan to recruit multiple cohorts to ensure we cover diverse life situations and to maximize what we learn:

- A general adult cohort of a few thousand people, broadly representative in age (say 20-85), gender, culture, etc. These people might be generally healthy, living ordinary lives. They give us a baseline for “normal” fairness dynamics (and let us see if even ordinary ups and downs show LoF patterns).
- A clinical cohort of a few hundred individuals with progressive illnesses (like early-stage neurodegenerative disease, metastatic cancer in remission, significant heart failure, etc.). These are folks for whom the horizon is foreseeably shortening, even if they might live years with treatment. Tracking them can directly test the end-of-life predictions in a naturally occurring way. We will work closely with their healthcare providers to ensure the study does not burden them and to integrate with their care plans.
- A caregiver cohort (perhaps overlapping with above): people who are caring for ill relatives. Caregivers often experience a unique oscillation of hope and despair tied to their loved one’s condition. Their perceived horizon can oscillate (“Will my loved one be here next year or not?”) and they may neglect their own needs until a crisis forces them to take relief. Studying them not only is ethically important (they deserve attention) but also provides a semi-experimental setting: when the patient’s condition worsens, the caregiver’s horizon shrinks (they fear time is short) and we can see if they start acting in QS-like ways (maybe reaching out for help or cherishing moments more).
- A late-life cohort (possibly 100-200 individuals) in partnership with hospices or assisted living facilities. These are people who are in that “closing chapter,” perhaps with a life expectancy under a year (though that’s always hard to say). We give them *maximal autonomy* – any participation is on their terms, and even small data from this group is incredibly valuable. They might choose to only do the self-reports and nothing else – that’s fine. The goal is to capture data in the *actual* final approach to end-of-life, to see if the neutrality and closure trends appear as predicted. Ethical safeguards here are strongest (consent re-checked frequently, etc., similar to Chapter 11’s outline).

We also build in some negative controls in the design. For example, we might randomly assign some participants to receive benign “wellness tips” or generic positive messages on their phone at random intervals. These should have no real effect according to LoF (they’re not targeted, just fluff), but they allow us to check for placebo or Hawthorne effects (people changing behavior because they know they’re observed). If mere participation or getting a generic tip changes something significantly, we have to account for that (maybe everyone in a study tries a bit harder at first). Ideally, those will wash out

or be negligible, confirming that any patterns we see are due to real life changes, not because we nudged them.

Now to the number-crunching side: Analytics from days to years. We have to bridge micro-scale fluctuations with macro-scale trends. We'll likely use a hierarchical modeling approach. At the lowest level, we have daily HCl estimates (with some uncertainty) for each person. We model each person's daily HCl as a latent state $x_t^{(i)}$ that evolves over time (like a random walk with drift plus adjustments when events happen). We effectively integrate the HCl to get an ongoing ledger $\hat{L}(t)$ for each person. Formally, we define the cumulative ledger as the time-integral of net felt experience. At the true theoretical level,

$$L(T) = \int_0^T F(t) dt$$

where $F(t)$ is the true net hedonic rate at time t . Our empirical estimate at time t is $\hat{L}(t) = \int_0^t HCl(\tau) d\tau$, i.e. the accumulated sum of our composite index (approximated in practice by a discrete sum at our sampling resolution). For readability below, we'll often write L_t for this estimated ledger $\hat{L}(t)$ (with its uncertainty implied), and treat $t = 0$ as the start of telemetry (or the earliest baseline window we observe). By the end of life ($T =$ death of mind), $L(T)$ is the final ledger total. The Law of Fairness posits $L(T) = 0$ for each, so we will test whether $\hat{L}(t)$ tends toward zero as $t \rightarrow T$ (within a defined confidence band) for our participants who reach end-of-life during the study.

We will also examine admissible-set proxies in the long data: things like daily “menu entropy” mentioned, or *action stickiness* (if a person does a reparative act one day, what’s the probability they do a similar act the next day?), *social graph tilt* (the proportion of interactions that are with close family or supportive people vs. casual or adversarial contacts), and a simple *channel access index* (how many of the key support channels – pain control, social support, etc. – are open on a given day). We model these as functions of the person’s current horizon H_t (which could be subjective or inferred from context like age/illness) and their ledger position $\hat{L}(t)$. For example, we hypothesize that as H_t shortens or $\hat{L}(t)$ becomes more negative, daily menu entropy E_t will decline (focus on fewer activities) and “repair stickiness” $S_{t, repair}$ will increase (once they start doing something to fix life, they keep at it). We’ll use mixed-effects regression, allowing each person to have their own baseline levels (random intercepts) and maybe their own sensitivity (random slopes) because not everyone will respond equally. We also include a wealth of nuisance covariates in these models: day of week (people have different patterns on weekends vs. weekdays), seasonal effects (mood can dip in winter, etc.), whether they were recently hospitalized, any therapy ongoing, and so on. We are especially careful with count-based outcomes like number of distinct actions: as

mentioned, we'll use Poisson models with checks for overdispersion, switching to Negative Binomial with a log link if variance is inflated. This ensures that if we see a significant "menu tightening" effect, it's not a statistical artifact of using the wrong distribution (we won't, for instance, interpret random noise spikes as meaningful just because we assumed a perfect Poisson when life is messier).

Another analysis technique: event-centered analysis. We'll align data around major events (like a diagnosis or bereavement) to see pre-to-post changes in key measures. To strengthen causal inference, we use techniques like *synthetic controls* or matched comparisons. For example, if 50 people in our sample experience the loss of a spouse, we can compare each of them to a "matched" person in the sample of similar age, baseline happiness, etc., who didn't experience such a loss, to see how their trajectories diverge. This is akin to treating life events as natural experiments: does a shock cause a pattern consistent with LoF (like the person's mood dips then gradually, through perhaps increased support-seeking, comes back to baseline) whereas the control person maybe stays steady? This helps isolate the effect of the event from general time trends.

A big part of our analysis is also a model competition. We will fit a suite of rival models to the same longitudinal data. These rivals include: a set-point adaptation model (everyone has a happiness set-point they revert to, with some damping factor), a reward-maximization or reinforcement learning model (people just chase rewards and avoid pain with no balancing constraint, possibly augmented with habituation to rewards), a predictive coding model (people react to prediction errors, not an explicit fairness goal), and perhaps a "no law" null model (any apparent balance is coincidental or due to interventions). We'll see how well each model explains the observed data patterns. We expect the QS-infused model (with horizon and ledger terms) to uniquely explain things like horizon-contingent shifts or those end-of-life compressions that the others can't. But we won't claim that until we prove it. All model comparisons will use rigorous out-of-sample validation – e.g. splitting data, using WAIC (Widely Applicable Information Criterion), leave-one-out cross-validation, predictive log-likelihood on held-out sequences. This prevents us from just fitting noise. If a simpler rival explains just as well as our model, we'll acknowledge that and refine or drop the claim accordingly. We essentially set up a "bake-off": whichever model predicts best (especially on those signature patterns) wins the explanatory trophy.

Long studies face inevitable challenges like missing data and drift. We plan proactively for these. People will miss surveys or sensors will fail; that's okay. We employ missing data modeling—for instance, using joint models where the probability of missingness can depend on someone's latent state (to handle if, say, people tend to skip reports when

they feel very bad, which is common). We also do sensitivity analyses: assume worst-case scenarios for missing stretches (e.g. maybe all missing days were terrible days) to see if conclusions still hold. For device drift (like if a wearable's step count calibration changes with a firmware update), we schedule periodic calibrations – e.g. we might ask participants to do a 2-minute walk test every 6 months to recalibrate step counts, or have known breathing exercises for HRV calibration. For participant drift (the idea that what “a 7/10 mood” means to you might change after a major life change or as years go by), we incorporate annual re-baselining sessions. In these, we might do a guided mood induction or reflective exercise to sort of “anchor” their scales again, or adjust scoring if needed. Essentially, we don’t assume static measurement – we continuously validate that our instruments measure the same constructs over time (a concept called measurement invariance). This is important so that any trends we see are real changes in people, not artifacts of shifting interpretation.

Ethics at this scale are just as crucial. We adopt a philosophy of “consent that travels.” This means participants can adjust what data they share at any time (dial it up or down) without feeling like they must drop out entirely. For example, someone might start comfortable sharing location data, then later feel uneasy – they can turn it off and continue in the study with other data streams, no problem. We also periodically (say every 6 months) do a mini re-consent, reminding participants what data of theirs we have and asking if they’re still okay with it. Privacy is built in: we try to keep raw data on the participant’s device whenever possible and only pull derived metrics (like “steps = 5000” rather than raw accelerometer readings). All data is encrypted and stored de-identified with a random ID. Even our research team might not see personal raw info – for instance, journal entries might be analyzed by an algorithm on the phone that just outputs topics or sentiment, rather than humans reading them unless the participant explicitly shares. We have an independent ethics board overseeing things, and *no data will be used for anything except this research* (no sharing with insurers or companies, obviously).

We stick to a simple rule: do no harm. We add no interventions that could increase suffering. For instance, any “horizon framing” we do (like asking someone to think about their future) is done very carefully and only if it’s emotionally neutral or positive – never to distress. If we ever prompt reflection (like end-of-life planning thoughts for older participants), it’s in a gentle, optional manner, often with supportive context. Late-life participants are always in control; they can choose minimal data collection (like just wear a device and no questions, or vice versa). We also provide *supportive tools* to everyone equally, not as experiments but as a benefit: e.g. an app feature to track their sleep or mood and give them feedback, or to record “legacy messages” for family – these

are made available not as a manipulation but as a thank-you for participating, and everyone can use them if they want, regardless of their data.

Alright, with all this machinery in place, what would count as success or failure for LoF in the long data? We predefine clear criteria. Supportive evidence would be seeing the family of patterns mentioned (ledger downturns leading to shifts toward relief behaviors, horizon-shortening leading to measurable changes in priorities, end-phase variance compression) appear *statistically significantly* and consistently across participants. For instance, if we find that in 70% of individuals, their mood variance in the last 3 months of life is at least 20% lower than their variance 6 months prior (and none of our controls can explain that), that strongly supports the compression idea. Or if, across the cohort, higher negative cumulative ledger values predict higher likelihood of engaging in help-seeking or restorative activities (with appropriate lags and controls), that supports the compensation idea.

Conversely, failure would be if these patterns do not emerge or if rival models explain them away. For example, maybe we find that people's happiness does tend to revert to baseline and nothing more – and any small horizon effects we thought we saw could be fully explained by, say, age or other factors. If after a couple of years we see no sign that people's behavior systematically changes in the way LoF predicts (and our sample and methods are robust), then LoF might simply be incorrect. Especially damning would be if some people's lives end with clearly unbalanced ledgers despite ideal conditions (we have to define what constitutes "clearly unbalanced" in data terms – e.g. final cumulative HCI far from zero with no balancing trend).

The chapter will detail these outcomes. But either way, by attempting this, we push knowledge forward. If LoF patterns appear across diverse lives tracked in this manner, it will be some of the strongest evidence in favor of a fundamental fairness principle in nature. If they do not, we will have amassed an unprecedented longitudinal dataset that will likely reveal other truths about well-being and adaptation, and we'll be in a position to say, "We tested the Law of Fairness in the real world, and here's how it failed," which is just as valuable scientifically.

Since it's crucial, let's explicitly state our Fail Conditions for this long view study (even if it's embedded in one of the sections): We would consider LoF unsupported in this domain if, for example, a large fraction of participants who died during the study had final-week average mood well outside the neutral band or significantly negative trends with no balancing – *despite* having good care (that hits RF1 from Chapter 11 in a longitudinal way). Or if none of the horizon or ledger-based predictors significantly improve our models of behavior over simpler models – meaning knowing someone's

cumulative pain/pleasure or perceived time left doesn't actually add any power to predicting what they do next. Also, if an adaptation model (which lacks LoF) predicts the longitudinal data just as well as the LoF-augmented model, then we haven't proven a need for LoF. Observing even one strong Fail pattern might not kill the law (maybe that person had special circumstances), but observing many, consistently (e.g. lots of people's lives end unbalanced *and* our models can't capture anything beyond ordinary adaptation) would be decisive against it. We intend to be totally transparent with these outcomes. We begin by tackling the foundational challenge: measuring the ebb and flow of a life in a reliable way.

What you'll get from this Chapter:

- A vision of “lifelong” experiments: See how modern technology allows us to study human lives in their full context, not just in lab snippets. This chapter will give you a concrete sense of what a multi-year human study looks like – from smartphone apps pinging you for mood ratings to wearables quietly logging your nights – and how we can knit that together into meaningful science. It’s like taking you on a tour of the future of psychological research, one where $N = 1$ (you) can still yield general laws when repeated across many.
- The signature patterns of fairness in the wild: You’ll learn the specific fingerprints we’re looking for in long-term data. This includes things like: how we’d recognize a “balancing act” in someone’s life trajectory, what a horizon effect looks like in everyday behavior (not just theory), and why a truly balanced final ledger implies certain measurable trends as the end nears. By the end, you could look at a hypothetical person’s multi-year happiness graph and have an idea of what we’d interpret from it in LoF terms.
- An appreciation for complexity (and how to tame it): Following our description of the statistical modeling, you’ll gain insight into how researchers extract clear signals from the noise of real life. We explain in approachable terms concepts like state-space modeling, why we use Bayesian hierarchical models, how we compare models with cross-validation, and so on – demystifying them so you understand *why* these tools are needed. This also underscores the point: we’re not just eyeballing squiggly lines and proclaiming victory; we’re doing rigorous, unbiased analyses to see if any effect is real.
- How we ensure participants remain people, not data points: The chapter devotes significant attention to the ethics and participant experience. You’ll see the safeguards and the philosophy behind them – essentially a blueprint for humane

long-term research. This not only assures you that our evidence (if we get it) wasn't obtained at the cost of exploitation, but it also shows how science can be done in partnership with participants. You'll likely take away some trust that *if* a law of fairness exists, we're going to find it without violating the fairness and dignity of those who help us find it.

- A head-to-head comparison with alternative explanations: As we lay out the rival models and control tests, you will see exactly how LoF stands apart from existing theories like the hedonic treadmill or simple homeostasis. We articulate what each rival would predict in our longitudinal study and then how LoF's predictions differ. This essentially gives you a mini-tutorial in the landscape of well-being theories and why LoF is a bold departure. By the end, you'll understand what would count as "just normal adaptation" in our data versus what would truly be evidence of something new (LoF).
- The endgame: how we'll know if the Law holds or folds: Finally, we make it very clear what outcomes of this years-long research effort would confirm the Law of Fairness and what outcomes would refute it. You'll know the exact benchmarks (like the ± 0.15 SD neutrality band, etc.) and combinations of findings that we've set as the bar for success. And you'll see that we haven't built a vague theory that can squirm out of failure – we've tied it down to empirical mastheads. If life doesn't obey this law, our long view study will be the one to loudly say so.

Subsections in this Chapter:

- **13.1 Longitudinal HCI, Practically** – Describes how we construct and validate the Hedonic Composite Index for long-term use. We explain the components (questions, sensor inputs) and how we weight or adapt them for individuals. This section also covers initial pilot testing of the app and sensors – e.g. ensuring people find the prompts acceptable, checking that the HCl correlates with known life events in a sensible way (validation). Essentially, it's the "methods" section for measurement.
- **13.2 Compression Near the End** – Focuses on one of the hallmark predictions: the affect variance compression as horizon goes to zero. We explain how we'll detect this in data (e.g. looking at rolling windows of variance for each participant as they age or as illness progresses) and differentiate it from other causes (like sedation or disengagement). We may present a hypothetical or real case study of someone in our pilot who entered hospice and show how their mood variability and activities changed. This section underscores why this is a crucial test and how exactly it will be measured.

- **13.3 Life Events as Ledger Shocks** – Details how we use major life changes to test LoF. We outline a few key event types (loss of loved one, major illness, major positive events like winning something) and what LoF would predict in each scenario. Then we describe our plan to compare trajectories around events, possibly including a visual example like “average mood trajectory 6 months before and after widowhood” from retrospective data or literature to set expectations. We clarify how we’ll use matching or synthetic control to attribute changes to the event. This section basically sets up life’s surprises as our unscheduled experiments and how we’ll learn from them.
- **13.4 Research Notes: Missingness and Hierarchical Models** – Gets into the weeds of the statistical approach. We outline the state-space model for mood and ledger, likely with an equation or two for the technically inclined (but explained in words as well). We discuss how we handle missing data formally – e.g. using Bayesian data augmentation or dedicated submodels for dropouts. We also explain the hierarchical nature: each person’s data is its own time-series, but parameters can be shared or partially pooled across people (shrinkage). If there are any priors or hyperparameters, we mention how we choose them (maybe based on earlier chapters’ results or plausible ranges). Essentially, we reassure the quantitatively-minded reader that we’ve thought this through and have solid plans to eke signal from noise.
- **13.5 Citizen Science Without the Creepiness** – This section likely addresses how participants engage with the study and even benefit from it. It might talk about features in the app like personal charts they can view, or a monthly summary email they get. We frame it as a form of “citizen science” – participants contributing to research and learning about themselves in the process. We emphasize transparency: participants will know what we’re measuring and why. Also, any aggregate findings will be shared with them first (or even co-interpreted with a participant advisory panel). The phrase “without the creepiness” signals that unlike big tech companies that collect data stealthily to manipulate behavior, our approach is openly collaborative and governed by ethics.

Where we go next:

In Section 13.1, we delve into how a person’s multifaceted daily experience can be distilled into a single composite index without losing the plot. This is the backbone of our long-term study – get this wrong and everything wobbles, get it right and we’re ready to chase LoF across years. Let’s see how we build that backbone.

13.1 Longitudinal HCI, Practically

This section turns the Hedonic Composite Index from a concept into a doable, ethical multi-year program that any serious lab or consortium could run. The goal is to estimate each participant's daily affective state and cumulative ledger $\bar{L}(t) = \int_0^t \text{HCI}(\tau) d\tau$, while tracking horizon perception H_t and admissible-set proxies (menu entropy, stickiness, channel access). Everything is designed to be low-burden, privacy-first, and preregistered.

13.1.1 Cohorts, cadence, and burden

Cohorts: We enroll multiple groups as described in the summary. The general adult cohort (on the order of thousands of participants) captures broad life variation. A clinical cohort (hundreds) focuses on those with known shortening horizons due to illness. A caregiving cohort observes people whose psychological horizon may fluctuate with someone else's health. A hospice-affiliated cohort covers late-life closure under careful ethical oversight. Each cohort has its own protocol adjustments (e.g. different prompt phrasing for older adults, special training for clinical groups), but all feed into the same core analysis.

Cadence: Participants in the main study receive EMA prompts throughout the day (morning, optionally midday, and evening) at random or scheduled times depending on preference. Passive data (sensors, phone logs) are collected continuously with periodic uploads. Lab or calibration activities (like yearly fMRI or monthly EEG headband sessions) are scheduled with plenty of flexibility and are entirely optional, especially for vulnerable participants.

Burden management: We set a strict guardrail that the median participant spends < 4 minutes per day actively engaged (answering surveys or tasks). Participants can “snooze” any data stream temporarily (consent travels with them), meaning they control the flow of data at all times. By design, missing some data is okay (our models account for it), so people are not pressured to respond when it's inconvenient or distressing.

13.1.2 The HCI inputs and minimal viable battery

To construct the Hedonic Composite Index (HCI), we gather several types of inputs with minimal intrusion:

A. Self-report (primary, but not exclusive): Each EMA survey includes:

Valence (-100 to +100): “Right now, how pleasant or unpleasant do you feel?” (slider from very unpleasant to very pleasant).

Arousal (0 to 100): “How activated or calm do you feel?” (slider from very calm to very activated).

Pain/Discomfort (0 to 10): Quick slider for physical or emotional discomfort level.

Horizon pulse (0 to 100): “Today, how far ahead does your life feel?” (slider anchored with descriptions like “just hours” up to “many years”).

Channel access checklist (yes/no for each): Are the following support channels available today? (Adequate sleep possible, effective analgesia available if needed, social support reachable, quiet/privacy available, financial breathing room). This checklist gauges which relief/recovery channels are open at that time.

B. Passive behavior (phone and wearables, opt-in):

Activity and sleep: Steps count per day; sleep onset/offset times and a sleep regularity index; a REM sleep proxy (if wearable provides it).

Communication graph: Number of calls/texts and unique contacts, reciprocity of communication (e.g. do others respond), average response latency, diversity of contacts (all based on metadata, never content).

Task cadence: Number of to-do items or calendar events completed (if user links a task app or calendar) and the time between completions – indicating pacing of goal-oriented behavior (no content of tasks, just timestamps).

Ambient mobility: Location variance (radius of gyration) and routine stability (how predictable daily mobility is) – derived from coarse location data or step patterns (we avoid GPS trails to protect privacy).

C. Physiology:

Cardiac and electrodermal: Resting heart rate and heart-rate variability (e.g. RMSSD from a morning reading), skin conductance (EDA) if the device supports it, and skin temperature. We include a weekly 60-second paced breathing exercise to capture a standardized HRV snapshot (deep breaths help measure parasympathetic tone).

Other wearables: If available, blood oxygen or other health metrics can be logged, but these are secondary.

D. Dreams (monthly focus week): One week per month, participants are asked to record any dreams they recall upon waking (via voice memo or a quick text description in the app). Those reports are later processed: trained coders (blinded to the person’s identity and context) rate each dream’s emotional valence (positive/negative) and look for

themes of threat, safety, or social reconciliation. We also use automated text analysis or embeddings as secondary features (looking, for example, at similarity to “reunion” or “danger” themes). Dreams provide a unique window into subconscious processing that might counterbalance daytime experiences.

E. Neural anchors (sparse):

Home EEG: Participants in upper tiers can wear a simple EEG headband at home on a quarterly or semiannual basis to perform brief tasks. For example, an auditory oddball task (hear a sequence of tones, occasionally an odd one out, measuring P300 amplitude/latency as an index of surprise processing) and a Go/No-Go task (measuring N2/P3 components for inhibitory control). We extract features like frontal theta bursts or P300 latency which tie to cognitive control and update processes relevant to QS.

Lab fMRI: A subset of willing participants (perhaps Tier C “methods helpers”) come in annually for a short fMRI session. The task might involve making choices between “repair” vs. “indulgence” options that have equal short-term utility, under varying horizon framings. We analyze predefined brain regions (ROI beta weights in right inferior frontal gyrus, dorsal ACC, ventromedial PFC, anterior insula, etc.) to see if the neural signatures of QS (e.g. control regions suppressing low-compensability choices) appear. These neural data serve to calibrate our HCI latent model and to verify that known control circuits behave as expected under horizon changes.

13.1.3 The HCI latent: model and outputs

We treat HCI not as a single measure but as a latent state that we infer from all the above indicators. A practical hierarchical state-space model can be defined as follows:

State model (affect dynamics): For person i , let $x_t^{(i)}$ be the latent true affect at day t . We model it as $x_t^{(i)} = \alpha^{(i)} + \phi x_{t-1}^{(i)} + \gamma \cdot \text{Ctx}_t^{(i)} + \eta_t^{(i)}$, where $\alpha^{(i)}$ is person i 's baseline affect level, ϕ captures inertia (how much yesterday's state carries into today; typically constrained to $|\phi| < 1$ for stability), Ctx_t represents contextual predictors (e.g. day of week, season, recent major events), and η_t is process noise. Essentially, affect drifts with some stability ϕ and reacts to known contextual factors.

We then assume our observed HCI composite HCl_t is a noisy reflection of the latent state x_t (with units in standardized HCU per day): $\text{HCl}_t^{(i)} \sim \mathcal{N}(x_t^{(i)}, \sigma^2_{\text{meas}})$, meaning the measured HCI (aggregating self-report, etc.) is normally distributed around the latent true state with some measurement variance σ^2_{meas} . This is a simplifying assumption; in practice our measurements are not strictly Gaussian, but we can transform scales or

use an appropriate likelihood (e.g. ordinal for Likert items, etc.) – the key is that they inform x_t .

Observation model (multi-channel indicators): We actually have multiple observed streams $y_{t,k}^{(i)}$ (e.g. self-report mood, HRV, sleep quality, behavior indices). We link each indicator k to the latent state: $y_{t,k}^{(i)} = \lambda_k x_t^{(i)} + b_k^T \text{Nuis}_t^{(i)} + \epsilon_{t,k}^{(i)}$. Here λ_k is the loading of latent affect onto indicator k (how strongly that channel reflects the latent HCl), and Nuis_t are nuisance covariates for that channel (e.g. motion artifacts for wearables, or time-of-day effects for self-report). $\epsilon_{t,k}$ is measurement noise for channel k . For example, a high-level model might say: today's self-reported valence = $\lambda_{\text{EMA}} \cdot x_t + (\text{maybe a bias term} + \text{noise})$; today's HRV = $\lambda_{\text{HRV}} \cdot x_t + (\text{controls for posture, time, etc.}) + \text{noise}$; and so on. By estimating λ_k , we combine channels optimally without any one channel dominating.

This hierarchical model is fit jointly so that we obtain outputs for each person: a daily estimated HCl $\hat{x}_t^{(i)}$ (with uncertainty bands) and from it a daily cumulative ledger $\hat{L}_t^{(i)} = \int_0^t \text{HCl}_\tau^{(i)} d\tau$ (with propagated uncertainty). We also compute the derived menu proxies (entropy, stickiness, etc. described next) for each day. In essence, for each participant we end up with a time series of their latent well-being and a running total of net experience.

Measurement invariance: In building this model, we assume that the relationship between latent affect and indicators is at least metric invariant across individuals (i.e. λ_k is generally shared or varies only within known bounds). We already handled calibration tasks (Section 13.1.5) to reduce individual differences. If only configural invariance holds (everyone has the same pattern of relationships but different scales), we may allow person-specific loadings but then limit comparisons to within-person changes. If the data and sample size allow, we strive for scalar invariance (equal intercepts) so that absolute levels of \hat{x}_t are comparable across people, but if not, we interpret results primarily within-person or within-group to avoid any bias from scale differences. In short, if cross-person measurement invariance is in doubt, we default to analyzing changes relative to each person's baseline rather than raw scores.

13.1.4 Admissible-set proxies you can compute

From the rich telemetry, we derive four key proxies that reflect the admissible set of actions (what the QS “lets” the person do readily) and how it changes:

Menu entropy E_t : the diversity of activities a person engages in during a day. We categorize each logged action (work, leisure, self-care, social, etc.) and compute Shannon entropy over these categories for day t . Higher entropy means a broad menu (many different types of actions); lower entropy means a narrow focus. QS predicts

entropy decreases when H_t is short or the ledger is in debt, as behavior funnels into certain channels (e.g. more time on meaningful or relief activities, less on miscellaneous ones).

Stickiness S_t (repair): the probability an action in a given category continues the next day. We focus on *reparative/relief* categories (e.g. spending time with family, engaging in therapy or healthful rest). We compute, for example, if a participant made a reconciliatory phone call or did a meditation session today, what is the chance they do something in that same category tomorrow? That is $S_{t, \text{repair}}$. QS predicts that as horizons shrink or ledgers go negative, $S_{t, \text{repair}}$ rises – once a compensatory activity starts, it sticks around – whereas stickiness for indulgent or trivial activities might not increase.

Social tilt T_t : the fraction of one's communication directed to close, supportive contacts vs. others. We label each call or message by relationship (participants can categorize certain contacts as “family,” “friend,” “work,” etc.). T_t could be defined as, say, proportion of outgoing communications that day that were with *kin/care* alters (family, close friends, therapists) as opposed to neutral or adversarial contacts. A rise in T_t indicates focusing one's social energy on supportive relationships. QS predicts T_t increases when H is short or ledger is low (people instinctively concentrate on those who matter most for repair or closure).

Channel access index A_t : the number of major support channels open on day t (out of the ones we track, e.g. sleep, analgesia, social support, quiet time). This is basically a count of “opportunities for relief/repair available” from the daily checklist. Interventions (like better pain management or providing respite care) should raise A_t . We expect QS effects to be stronger (and ledgers more balanced) when A_t is high, because the system has means to compensate. We will use A_t as a moderator in analyses (e.g. does horizon shortening lead to variance compression only when channels are open? It should, if LoF is genuine).

All four indices are standardized within-person and then analyzed primarily as outcomes or predictors in relation to \hat{H}_t and L_t . *For instance, in models we might include terms like \hat{H}_t^{-1} (the inverse of horizon length, capturing “closeness to the end”) and $|L_t|$ (absolute ledger imbalance) to predict E_t , $S_{t, \text{repair}}$, etc., controlling for day-level nuisances.* By standardizing within person, we focus on changes relative to one's typical behavior, which is safer if people have different baseline entropy or stickiness levels.

(Note: as mentioned in the Summary, any counts involved in these proxies are modeled with appropriate distributions. For example, if we analyze the raw count of open channels

A_t or number of unique alters in communication, we treat them as Poisson variables initially, check dispersion (if the variance exceeds the mean notably, >1.2×, we use a Negative Binomial), and we always report which link function was used. This prevents spurious findings due to model misfit.)

13.1.5 Monthly calibration “mini-lab”

To keep measurements calibrated and honest across a long timespan, the system includes a brief monthly in-app calibration session (opt-in, but encouraged). In about 10 minutes, the participant goes through a few standardized tasks:

Affective pictures: The app shows a set of standard images (akin to IAPS, International Affective Picture System) that span a range of pleasant/unpleasant and arousing/calming content. The participant rates their feelings for each. This helps align the self-report scale over time (we can detect if someone’s use of the 0–100 valence scale drifts, and correct for it).

Paced breathing and orthostatic HRV: The participant does 1 minute of slow breathing and perhaps a quick sit-to-stand test while the wearable captures heart signals. This provides a controlled HRV measure and checks if the device’s readings are consistent (e.g., if firmware updates changed HRV calculation, we’d notice via this standard task).

Cognitive control task: A 2-minute Go/No-Go or Stroop-like game on the phone to gauge any drift in inhibitory control or attention baseline. Over years, if someone’s cognitive capacity changes (due to aging or illness), it could affect how QS operates (since QS partly relies on cognitive control to suppress certain actions). Tracking this ensures we know if any loss of variance is just due to cognitive decline rather than QS per se.

Anchor vignettes for Φ : We present a few very short stories or scenarios and ask the participant how they’d likely respond or feel. These are designed to probe the features of *compensability* (our Φ model). For example, one vignette might describe a situation offering relief (ending a long struggle) versus pure pleasure; another might present a repair opportunity (making amends) versus a neutral option. By having participants react, we “refresh” our model of what counts as high- Φ (compensatory) actions for them. Essentially, we recalibrate the mapping of scenario features (ReliefGain, RepairGain, HarmRisk, OptionFlexibility, etc.) to their values. This helps keep the Φ feature model up-to-date if a person’s values or circumstances shift.

These mini-lab tasks, while optional, greatly improve data quality by ensuring our scales and inferences don’t quietly drift over long periods.

13.1.6 Data dictionary (minimal viable fields)

For transparency and replication, we define the minimal data fields we record:

Daily core data: Date/timestamp; EMA responses (valence, arousal, pain scores); horizon slider value; channel checklist responses; sleep duration and a sleep regularity metric; resting HR and HRV; step count; derived menu entropy for the day; repair stickiness; social tilt; channel access index A_t ; and our estimates for HCl (\hat{x}_t) and ledger L_t (with their credible intervals).

Weekly entries: Text of the reflective weekly journal (tokenized or summarized on-device to preserve privacy) and any tagged life events that week (e.g. “bereavement” or “job change” tags, entered by participant or via linking to a calendar).

Monthly: Calibration task results (ratings for standard images, HRV from breathing test, etc.) and device quality control flags (e.g. if a wearable’s data quality was low, firmware updates, etc.).

Quarterly: EEG metrics from home headband (P300 amplitude/latency, N2 No-Go amplitude, frontal theta power, etc.).

Annual: fMRI ROI values (if done) capturing contrasts related to $\Phi \times H^{-1}$ interactions (for example, increased activation in control regions when choosing a low- Φ action under short horizon).

This data dictionary guides what we preregister and ensures we only collect what we truly need.

13.1.7 Missingness, drift, and quality control

Despite best efforts, missing data will happen. We incorporate specific strategies:

Planned missingness: We intentionally schedule occasional “skip days” for some participants (no prompts) to later assess if missingness is truly random or if people only skip on hard days. By having some researcher-imposed missingness, we can compare patterns.

Joint modeling of missingness: As detailed in 13.4, we model the probability of missing data as potentially dependent on the latent state. For example, if evening EMAs are more likely to be missed when someone is distressed (MNAR: Missing Not at Random), our joint model can detect that via a parameter γ linking latent affect to missingness (see Section 13.4.2). We also do delta-adjustment sensitivity analyses: assume, say, that all missing evening reports were actually 0.5 SD more negative than observed ones, and see if conclusions change (a “worst-case” scenario test).

Device QC: We log device metadata such as firmware version, battery level, and any sensor quality metrics (like proportion of time heart sensor had good contact). If a device's data quality drops (e.g., lots of motion artifact or data gaps), we flag those periods. In analysis, we can include such flags to avoid misinterpreting noisy data as true variance changes.

Participant check-ins: If someone's data suddenly changes (e.g., they stop responding entirely or their signals shift drastically), we have protocols to gently check in, ensuring they're okay and seeing if they need a break or support. Ethically, we treat sustained non-response not just as missing data but as a potential indicator of participant burden or distress.

13.1.8 Where we go next:

In summary, Section 13.1 laid out how we *practically* measure longitudinal HCI and ensure data integrity. With this foundation, we now turn to specific patterns the Law of Fairness predicts in these longitudinal data, starting with what it means to see compression near the end of life.

13.2 Compression Near the End

If the Law of Fairness is true, end-of-life is where its tightest constraint should appear. As subjective horizons shrink, the Queue System (QS) should tilt one's "menu" of actions toward reparative and relief-oriented choices, while suppressing pursuits that would create uncompensable pain or debt. At the population level, this implies a measurable pattern in the data: variance compression of the daily HCl (emotional ups and downs become muted) and a tilted composition of actions (more time on closing, healing activities, less on adversarial or trivial ones), with the cumulative ledger $L(T)$ drifting toward neutral as $t \rightarrow T$.

13.2.1 What "compression" means operationally

We define hedonic compression in concrete, testable terms. We track three families of within-person signatures (computed over sliding windows of, say, 3–4 weeks each):

Volatility drop: A decline in short-term variability of affect. We compute $\Delta\sigma^2_{HCl}(t)$, the change in HCl variance between successive windows (e.g. compare variance in the past 3 weeks vs. the 3 weeks before that). Formally, $\Delta\sigma^2_{HCl}(t) = \sigma^2(HCl_{[t-w:t]}) - \sigma^2(HCl_{[t-2w:t-w]})$ for window length w . The prediction is $\Delta\sigma^2_{HCl}(t) < 0$ as the clinically estimated horizon H_t shrinks. In plain terms, as someone enters their final phase, their day-to-day mood swings should get smaller (variance in recent days lower than a month ago).

Skew normalization and drift toward neutral: We look at the distribution of recent HCl values and ledger changes. We expect the skewness of the HCl distribution to approach zero (no strong skew toward only positive or negative extremes) and the absolute change in ledger over a window, $|\Delta L_w(t)| = |L_t - L_{t-w}|$, to shrink. This would indicate the stream is leveling out around neutral rather than taking wild swings. So as the end nears, recent experiences are neither highly skewed good nor bad, and net change is modest.

Menu tilt and stickiness: We examine behavior composition:

Daily entropy E_t (diversity of activities) is expected to drop as certain activities fall away.

Repair/relief stickiness $S_{t,repair}$ should rise – if the person engages in something like reconciling or comfort-seeking, they continue in that vein.

Social tilt T_t should shift toward kin/care contacts.

These behavioral shifts serve as *external validators* of compression: it's not just that mood swings less, but also the person's focus narrows to a core set of meaningful or

soothing activities. We will test all these as functions of horizon H_t and current ledger position L_t , using within-person mixed models with rich controls.

13.2.2 Distinguishing real compression from artifacts

One challenge is that apparent “peacefulness” at end-of-life could be an *artifact* of other processes, like heavy sedation or physical fatigue, or simply because measurement tools hit a ceiling/floor. We therefore deploy layered discriminators to ensure any observed compression is genuinely due to QS dynamics (closure) and not a false positive:

Sedation vs. closure: If someone is heavily medicated (e.g. morphine in hospice), their arousal and affect may flatline – that’s sedation, not necessarily QS. We differentiate by signature: Sedation profile is characterized by a sharp drop in arousal (maybe near-zero activation), changes in HRV consistent with reduced arousal (medication and illness effects can be bidirectional), a global entropy collapse (all activities drop off equally), and crucially no selective increase in reparative actions (they’re not actively doing closure tasks, they’re just doing very little of anything). In contrast, QS closure profile might show only a modest decline in arousal (they might be calmer, but not comatose), a selective entropy drop (certain categories of activity—especially trivial or avoidant ones—shrink more than others), and evidence from neural anchors: e.g. in an EEG or fMRI, we might see rIFG/ACC (often implicated in inhibitory control) engaged during suppression of low-Φ options, and vmPFC showing stronger value-related signals for reparative options (interpreted as correlational support). In short, if it’s true closure-driven compression, the person is still cognitively *engaged* in a focused way, rather than universally dampened.

Fatigue vs. closure: Severe illness often brings fatigue, which can reduce activity across the board. How to tell apart fatigue (just low energy) from QS-driven narrowing? Fatigue would cause a broad performance decrement: the person does less of everything and might disengage socially, but without a particular tilt toward meaningful interactions. We’d see maybe more sleep, less of all types of activities, and no special preference for calling loved ones or finishing tasks. Closure, on the other hand, predicts that even if total activity drops, the *composition* shifts – e.g. the person might have fewer total awake hours but spends a greater *fraction* of those hours on, say, writing farewells or connecting with family. We specifically look for increased proportion of repair contacts and task finalizations (like completing one’s will or organizing affairs) even if overall action count declines.

Ceiling/floor effects on measurement: If our mood scales have a fixed range (say -100 to +100), someone’s affect might appear “stable” simply because they’ve hit an upper or

lower limit (can't report feeling worse than the bottom of the scale). To guard against this, we use rank-based volatility metrics like median absolute deviation, which are less sensitive to hard limits. We also incorporate Bayesian measurement models to estimate true variance beyond the scale limits. Additionally, we do dynamic range checks: thanks to our monthly calibration (13.1.5), we inject some standardized stimuli. If we notice that a participant's self-report range has collapsed (e.g. they only use 50–60 on the slider anymore) *but* their physiological or behavioral responses to calibration still show capacity for highs and lows, that's suspicious — it could mean they are stoically reporting a narrow range while still experiencing swings. We flag such cases as possible measurement compression artifacts.

By applying these discriminators, we ensure that when we say “variance compressed near the end,” we mean *specifically due to QS-style closure*, not simply because the person was sedated, exhausted, or because our measurements failed to capture variation.

13.2.3 Horizon estimation and staging

A core part of this analysis is knowing *how close to the end* someone is — i.e. their remaining horizon H_t . We use both subjective and clinical horizon measures and then combine them:

Subjective $H_t^{\{subj\}}$: This comes from that daily horizon EMA (“How far ahead does your life feel?”). We treat it as a rough numerical scale (0 meaning “I only see hours ahead” up to 100 meaning “years ahead”). While noisy and mood-influenced, it provides the person’s internal sense of future.

Clinical $H_t^{\{clin\}}$: For participants with terminal illness or very old age, clinicians often have some prognosis categories (like “months left” vs. “weeks left”). We encode broad bands (e.g. >6 months, 1–6 months, weeks, days) based on medical opinion or hospice staging. For healthy adults, this is essentially “unknown/long.”

We then fuse these (after mapping both to a common scale): $\hat{H}_t = w_{subj} \cdot H_t^{\{subj\}} + w_{clin} \cdot H_t^{\{clin\}} + \xi_t$, a weighted combination. The weights w_{subj} , w_{clin} are learned hierarchically (we might let the model find how much to trust people’s own horizon sense vs. clinical indicators). ξ_t is an error term. If they disagree (e.g. a patient feels hopeful, giving a long subjective horizon, but clinicians estimate weeks), we don’t throw out either – instead we flag those cases for qualitative adjudication, meaning we might examine that trajectory separately or ensure the analysis is robust to whichever measure is right. All analyses will be preregistered to be run with subjective horizon alone, clinical alone, and the fused measure, to see if results converge.

Accurate staging of horizon is vital: a person might not *know* when they are 6 months from death (unless in hospice), so subjective horizon is an imperfect proxy. Conversely, clinical predictions can be off. By combining them, we aim to get the best estimate possible for how “close to the end” each data point is, which is the key independent variable driving compression effects.

13.2.4 Primary compression model

With horizon in hand, we set up our primary statistical model to test compression. It is a within-person mixed-effects model over time. For each person i and time t , we model something like:

Outcome 1 (affect variance): $\sigma_{\text{HCl},t}^{(i)}$ (the standard deviation of that person’s HCl in a recent window) as a function of horizon and ledger:

$$\log \sigma_{\text{HCl},t}^{(i)} \sim \beta_0 + \beta_1 \hat{H}_t^{-1} + \beta_2 |L_t| + \beta_3^T \text{Nuis}_t + u_i + \epsilon_{it}.$$

Here \hat{H}_t^{-1} is the inverse horizon (shorter horizon = larger value), $|L_t|$ is the absolute ledger imbalance at time t , Nuis_t includes control covariates (e.g., whether in hospice, current pain level, medication, season), u_i is a person-specific random intercept, and ϵ_{it} is residual error.

We include person intercepts u_i to focus on within-person effects.

Outcome 2 (entropy): $E_t^{(i)}$ (daily menu entropy) similarly as $E_t^{(i)} \sim \gamma_0^{(i)} + \gamma_1 \hat{H}_t^{-1} + \gamma_2 |L_t| + \dots$, with analogous terms.

Outcome 3 (repair stickiness): $S_t, \text{repair}^{(i)}$ as $S_t, \text{repair}^{(i)} \sim \delta_0^{(i)} + \delta_1 \hat{H}_t^{-1} + \delta_2 |L_t| + \dots$,

We won’t write all terms “...” but they include similar nuisance controls and random effects structure for each outcome.

The QS predictions we’re testing in this model are: $\beta_1 < 0$ (shorter horizon \rightarrow smaller affect variance, so \hat{H}^{-1} has a negative coefficient meaning as horizon shrinks, σ goes down), $\gamma_1 < 0$ (shorter horizon \rightarrow lower entropy), and $\delta_1 > 0$ (shorter horizon \rightarrow higher repair stickiness). We also expect coefficients on $|L_t|$ to perhaps capture that when the ledger is very imbalanced, similar adjustments happen (even if horizon isn’t short, a big negative ledger might trigger a local compression as the system tries to recover).

All these are *after controlling for* factors like seasonality (maybe winters lower variance anyway), medications (pain meds might lower variance, we account for that), sleep debt, hospital days, device changes, etc. We will look at the estimates and confidence intervals of $\beta_1, \gamma_1, \delta_1$ primarily.

Statistically, we might implement this with Bayesian hierarchical models or frequentist mixed models. Either way, we will report not just significance but effect sizes: e.g. does within-person HCl variance drop by, say, 0.15 SD (15%) as horizons enter the final month? We have a *pre-registered threshold*: compression would require at least on the order of mean change of 0.15 SD (± 0.15 z) in affect and slope on horizon of ± 0.05 z per day (if measuring vs. days to live for terminal cohort), and a variance ratio ≤ 0.80 (final phase variance less than 80% of earlier phase). These thresholds (Smallest Effect Size of Interest, SESOI) guard against over-claiming a trivial change as meaningful compression. We will treat the compression hypothesis as supported only if changes meet or exceed those magnitudes with high confidence.

(Technical note: Entropy and stickiness might not be normal — entropy is bounded [0, log K] and stickiness is a probability [0, 1]. We'll use appropriate links (e.g. beta or logistic models for stickiness, maybe a truncated normal or beta for entropy if needed). Count of actions, if directly modeled, uses Poisson/NB as noted. All such modeling choices are preregistered.)

13.2.5 Composition tests that cannot be faked by sedation

Beyond statistical models, we do a more intuitive composition analysis: what fraction of a person's day is spent in different categories as the end nears? We decompose each day's activity into three broad categories and examine their changing shares (using compositional data analysis techniques, like additive log-ratio transforms to properly handle the fact that fractions sum to 1):

Repair/relief activities: e.g. calling or visiting close family, making reconciliations or apologies, managing symptoms/pain relief, quietly resting or meditating – essentially actions that directly contribute to emotional closure or relief.

Neutral routine: e.g. routine administration, watching TV or casual reading, basic chores – activities that are neither highly compensatory nor particularly harmful, just daily life maintenance or idle pastimes.

Adversarial/avoidant activities: e.g. engaging in conflicts, “doomscrolling” negative news, risky indulgences, self-sabotaging acts – things that could add negative ledger entries or avoid dealing with the important issues.

QS predicts an increasing share of repair/relief and a decreasing share of adversarial/avoidant as $H \rightarrow 0$, beyond just a general decline in total activity. Sedation, by contrast, might reduce total activity (the pie gets smaller) but wouldn't selectively increase the “repair” slice – sedation would just shrink all slices uniformly. To test this, we examine compositional changes: for each person, in their last months, do the

fractions shift? For instance, if a patient used to spend 10% of time on social calls and 10% on worry scrolling, and near the end it's 20% on social calls and 0% on doomsscrolling (with overall less activity), that's a selective shift consistent with QS closure rather than sedation. We will quantify these with log-ratio analyses (e.g. the log of (repair%/adversarial%) should increase if closure is happening).

In short, we expect to see a rising share of repair/relief and a falling share of adversarial, even accounting for less overall activity, whereas pure sedation would show reduced activity *without* such selective changes.

13.2.6 Social graph tilt as an external validator

Another external indicator of closure is in the social network. Every day we can derive features of the participant's ego-network:

Degree: how many unique people they communicated with that day.

Assortativity: fraction of communications that are with close kin/care persons versus others.

Reciprocity: whether communications are mutual (messages/calls returned).

Latency: response time in conversations.

We also define a “repair index” in communications: proportion of communication that involves previously estranged contacts or explicitly revolves around closure themes. For instance, if our NLP on text messages (done on-device) flags that someone is talking about forgiveness or final wishes, that counts toward a closure topic (we don't see the content directly, just a topic label). Or if the person contacts someone they haven't spoken to in years (estranged sibling, etc.), that's a significant event.

Predictions: As end-of-life nears, we expect the social graph to contract toward family and close friends, meaning degree might drop but those who remain in contact are core supporters (increased kin assortativity). Reciprocity might actually increase because both the participant and their contacts prioritize responding (fewer dropped conversations). Latency likely shortens for important exchanges – e.g. if someone is making final arrangements or goodbyes, replies happen faster. These can be measured: average reply time to family vs. others, etc., and should show faster engagement on meaningful ties.

So, an external validator of compression is that even their social life shows a focusing: e.g. instead of occasionally arguing with an acquaintance on social media (adversarial contact), the person might drop that entirely and spend their limited energy calling their

children. We'll test these patterns by comparing social metrics in "near end-of-life" periods vs. earlier periods, controlling for things like number of friends alive, etc. If we see, say, increased proportion of kin contacts and more two-way exchanges near the end, that strongly supports QS's influence and not just a random withdrawal.

13.2.7 Ledger drift corridors and neutral approach

How do we formally test if life's ledger approaches neutrality by the end? We don't expect the ledger to hit exactly zero on any given day. Instead, we define corridors of acceptably "near-neutral" net experience over longer windows. For example, we might set a band $[-c, +c]$ HCU for cumulative change over 30 days, where c is chosen in the same ledger units as the 30-day ΔL (i.e., the 30-day sum/integral of daily HCl), based on a meaningful effect size. We pre-specify corridors for 30-day, 60-day, and 90-day changes.

Then we test: in the final stretch of life (say the last 90 days), is the person's ledger change more often inside that corridor compared to earlier in life or compared to matched others not near death? Concretely:

Define $\Delta L_w(t)$ as ledger change over window w ending at t . We choose windows (e.g. $w = 30, 60, 90$ days) and a tolerance c_w for each (predefined in HCU units and scaled to within-person baseline variance).

Compute the posterior probability that in the *last* three windows before death, ΔL stayed within $[-c_w, +c_w]$ more frequently than it does for a control group of similar age/diagnosis individuals observed in non-terminal periods.

If this probability is high (and higher than chance comparisons), we have evidence of "staying in the neutral corridor" as the horizon closes. Essentially, we're asking: as H^{-1} increases (horizon shrinks), does the ledger increasingly hug close to zero net change?

We call a result supportive if corridor occupancy indeed rises significantly with H^{-1} and the other composition signatures (variance down, repair up, adversarial down) co-occur at the same time. It's a stringent test: not only must variance shrink and actions tilt, but those must coincide with the ledger actually not straying far from zero near the end.

13.2.8 Blinds, adjudication, and dignity

Throughout this analysis, we implement measures to ensure bias doesn't creep in and dignity is preserved:

Blinded coding: Qualitative data like dream reports or participants' journal notes (if used for analysis) are coded by individuals who do not know how close that participant was to death at the time, or any outcomes. They might just get randomized sets of text to rate for

emotional tone or closure themes. This prevents coders from unintentionally confirming our hypothesis (e.g., not assuming a dream must be positive just because it's near the end).

Blinded clinical summary: When we use clinical notes (like hospice nurse observations), a separate team translates those into event codes or symptom scores. Analysts working on the core data only see these abstracted codes, not the narrative. This avoids any cherry-picking or storytelling that “oh, this person was clearly at peace” without evidence. We rely on coded data.

Participant dignity: As part of the study design, we offer participants private “legacy tools” (like the ability to compose messages to loved ones to be delivered later, or create digital memory albums) regardless of whether they are in a control or experimental condition. These are not manipulations but resources. We deliberately do not force any behavior for the sake of data. Importantly, we do *not* do anything like shorten someone’s perceived horizon artificially or dramatize their situation for a test (that would be unethical). All participants, especially in late-life, have full control and plenty of opt-outs. Our motto: widen channels, *never* constrict them, and never treat a person as merely a subject — they are a person living their life, and we measure from the sidelines.

13.2.9 Interim monitoring and early stopping

Because we’re dealing with potentially vulnerable populations and a long study, we set up interim analyses with stopping rules:

We identify three co-primary endpoints (based on our predictions): Horizon $H^{-1} \rightarrow$ HCl variance down (a significant negative association between inverse horizon and affect variance), $H^{-1} \rightarrow$ repair stickiness up, $H^{-1} \rightarrow$ adversarial share down.

We use an alpha-spending approach (e.g., O’Brien-Fleming boundaries or similar) to peek at the data periodically with strict thresholds for early success or futility. For instance, if by midpoint the horizon-variance effect is so strong that it meets the pre-set boundary, we might declare success on that endpoint early (if ethically appropriate to share).

Likewise, we have futility stopping: if it becomes clear that, say, adversarial share isn’t dropping at all with horizon in a large sample, we might stop testing that particular hypothesis or reduce resources on that measure.

We also maintain negative controls during monitoring: for example, we might introduce random “countdown” timers to some participants that have no real significance (explicitly framed as an arbitrary administrative date, not a life-horizon cue) to ensure that

if we see horizon effects they're truly tied to real horizon, not just any countdown. These negative controls should yield null results (no compression triggered by a fake horizon), otherwise something's wrong. We continuously verify that such placebo conditions remain null.

13.2.10 Worked vignette (what the pattern looks like)

To illustrate, imagine a concrete example:

“E.” is a 72-year-old participant with progressive heart failure. About six months before her death, doctors start to hint that time may be limited; her subjective horizon also begins to shorten (she stops planning far into the future). Over the next 90 days, we observe the following:

Affect variance halves. In the months prior, her daily HCl might fluctuate widely (some days +20, some -50, etc.). In the final three months, those swings dampen to perhaps half the range. Concurrently, her median HCl, which had been mildly negative, slowly rises toward zero (she reports less frequent very bad days).

Menu entropy drops. She cuts out certain activities (for instance, she quits a contentious community board that caused stress and scales back on watching upsetting news). She doesn't do as many different things in a day, focusing on a few core activities. Importantly, repair actions become “sticky.” Once she decides to reconnect with her estranged sibling and they start talking, this becomes a daily routine; she consistently engages in that meaningful interaction rather than one-off.

Social graph contracts to family. She communicates with fewer people overall, but a larger share of her communications are with close family. We see her reciprocity go up (she responds to every message from loved ones promptly) and average response latency shrink – those conversations happen in near-real-time now, whereas she ignores less important contacts.

Health and environment factors: During this period, she is *not* heavily sedated (pain is managed well enough without clouding her mind). Her sleep regularity actually improves a bit (she sets a routine of going to bed and waking up earlier, perhaps preparing for closure). In her final lab fMRI (done maybe ~3 months before passing), we even notice her vmPFC showing stronger activation for “reparative” choices compared to last year – suggesting her brain is prioritizing those.

All these signs point to the QS closure pattern rather than just the effects of illness. Her pattern does *not* match sedation (she was awake and making deliberate choices) nor simple fatigue (she *did* things, just selectively). It's as if her system “knew” to settle

things: emotional volatility calmed, negative ledger entries were avoided, and she gravitated toward things that gave peace or closure.

13.2.11 Fail conditions (what would count against QS)

We also pre-specify what findings would count as failures for the QS end-of-life hypothesis. Any of these observed robustly (especially if in combination) would seriously challenge LoF:

No variance compression: If as horizons shrink, we *do not* see a drop in affect variance – or worse, if variability stays high or even increases – *after* controlling for sedation, fatigue, and measurement issues, then the predicted constraint isn't showing up.

No composition tilt: If the share of repair/relief activities does not increase (stays flat or negligible) and the adversarial/avoidant share doesn't decrease (or even grows) as death nears, that contradicts QS. Essentially if people continue in the same proportions of activities or get more bitter/avoidant on average, that's a failure of the fairness mechanism.

No neutral drift: If our ledger corridor test finds that near-terminal periods are no more likely to be near-neutral than earlier periods – e.g. people still accumulate large net negatives or positives in their final weeks just as often as before – then LoF's key claim of final balance is unsupported.

Rival explanations suffice: If adaptation plus fatigue (for instance) can reproduce the entire pattern we see *without* needing any horizon or QS parameter, then QS might not be needed. For example, maybe everybody just adapts and gets tired at the end, and that combination flattens variance with no special fairness constraint. If we fit such rival models and they predict the data as well as our model including horizon terms, it counts as an against-LoF result.

Replication of Fail patterns: If the above Fail patterns are seen consistently in multiple cohorts or contexts, we would have to conclude that what we're seeing is just tendency or coincidence, not a law. The instruction in our falsification plan is that if two or more of these fail conditions replicate with high power, we “retreat to tendency accounts or reject” the LoF hypothesis for end-of-life balancing.

13.2.12 Ethics: the line we do not cross

Because this section deals with end-of-life, it's worth reiterating ethical boundaries clearly:

We do not and will never intentionally shorten anyone's horizon or make them feel time pressure for the sake of a study. We would never create suffering or panic just to see if compensation kicks in – that is off-limits.

We provide comfort and support (widening channels like better pain control, easier communication tools, better sleep conditions) equally to participants; none of our experiments involve depriving someone of relief to see what happens. We are measuring a natural process, not provoking it.

In late-life studies, participation is entirely opt-in and can be revoked at any time without any penalty or even explanation needed. If someone prefers not to log anything in their final days, that choice is paramount – dignity over data.

Our role is observation and modeling, not steering lives. We explicitly tell participants and ourselves that we're not trying to "optimize" their last chapter or test them in any way; we're simply there to respectfully learn, and our first priority is their comfort and autonomy.

Takeaway: Compression is not presented here as a romantic notion of peaceful endings – it is framed as a *quantitative, testable pattern*. Specifically, as horizons shrink and if channels for relief are open, we expect to see: variance down, repair-oriented actions up, adversarial actions down, and ledger drift bounded within a corridor. If those signatures selectively appear when they should (and not due to trivial explanations), it's exactly what LoF and QS predict at life's edge. If they fail to appear, LoF might be a false lead.

13.2.13 Where we go next:

Next, Section 13.3 operationalizes life events as ledger shocks – effectively using major events (good or bad) as natural experiments to further test admissible-set dynamics and compensation *in vivo*.

13.3 Life Events as Ledger Shocks

Some days are pebbles; a few are boulders dropped into the stream of life. Marriage, bereavement, a lottery win, trauma, a big promotion, a disabling injury – these major life events jolt the hedonic state and alter the menu of possible actions far more than ordinary day-to-day fluctuations. This section treats such major life events as exogenous shocks to a person's ledger (and possibly to their horizon) and shows how we can leverage them to test the Law of Fairness (LoF) and the Queue System (QS) without breaching ethical boundaries. In essence, life itself performs experiments by delivering shocks; our job is to measure the before-and-after.

13.3.1 What we mean by a “ledger shock”

First, let's define the term. Recall HCl (the measured daily hedonic state) and

$L_t = \int_0^t HCl(\tau) d\tau$ (the cumulative ledger up to time t). A ledger shock is an event E at some time t_0 that produces either a discrete jump in the hedonic state HCl_t , or a sustained change in the slope of L_t (i.e., the trend of hedonic experience), beyond what our baseline model would predict for that time.

Essentially, it's when reality deviates sharply from the forecasted well-being trajectory due to something happening. We can formalize an impact function: $\Delta E(t) = HCl_t - HCl_t^{\{counterfactual\}}$, where the counterfactual is an estimate of what HCl would have been at time t had the event E not occurred. We estimate that using matched comparisons or pre-event trends (more on methods soon).

QS expectation: For negative shocks (losses, trauma, etc.), we expect the QS to respond by tilting menus toward relief and repair. That might manifest as, after a tragedy, the person becomes more likely to seek support, comfort, or restorative activities and avoids new risky or harmful endeavors. The stickiness of those reparative actions should increase as well (if they start going to a support group, they keep at it). For positive shocks (windfalls, big successes), QS should prevent unbalanced indulgence: we expect a tendency toward meaningful expenditure of that good fortune or outward care (for instance, a windfall might lead someone to share with family or donate, rather than purely self-indulge endlessly). This especially holds if horizons are short – a person late in life who gets a surprise award might channel it into legacy or care activities rather than a frivolous spree. In summary, QS implies *compensatory adjustments* following shocks: big negatives prompt counteracting positives (not just eventual adaptation to baseline, but perhaps overshooting above baseline to compensate), and big positives prompt grounding or even protective negatives (not punitive, but stabilizing to avoid overshoot).

We will test these nuanced predictions against alternatives like “regression to the mean” or standard adaptation (which would predict after a shock, people eventually just go back to baseline with no overshoot and no particular tilts in behavior composition).

13.3.2 Taxonomy of events we can study now

We categorize major life events into classes, because each might have different dynamics:

Loss/Threat events: e.g. death of a close loved one (bereavement), a breakup/divorce, being laid off or fired, eviction or foreclosure, receiving a major medical diagnosis (like cancer), or experiencing a natural disaster or violent incident. These are negative shocks that often acutely drop HCl.

Gain/Opportunity events: e.g. a job promotion, getting married or committing to a partner, the birth of a child, receiving an inheritance or lottery win, winning a major award or recognition. These are positive shocks that spike HCl upward (at least initially).

Relocation/Regime change: e.g. moving to a new city or country, migration (especially if forced or for refuge), retirement (exit from a long-term role), release from prison or completing military service. These are events that can be mixed (some positive liberation, some stress of change).

Care transitions: e.g. starting to care for an ailing relative or conversely the end of a caregiving role (perhaps the person you cared for passed away or recovered), enrolling in hospice, etc. These alter one’s responsibilities and horizon perspective.

For each category, we pre-specify inclusion rules. Importantly, we rely on objective, time-stamped definitions: e.g. bereavement is logged by the date of death certificate of the loved one; job loss by the date on a termination letter or last pay stub; marriage by the wedding date, etc. We do this to avoid any cherry-picking or hindsight bias in saying when the “event” happened. The time t_0 of the event is fixed based on external documentation, not based on when we notice a mood change. This ensures we don’t, for example, slide the window to make things look compensatory after the fact.

Also, each event type is considered separate in analysis because the patterns might differ (bereavement likely has a different time-course than, say, marriage). We’ll analyze each category on its own as well as potentially aggregating similar “negative shocks” vs. “positive shocks.”

13.3.3 Design 1: event-study panel (difference-in-differences)

One analytical design is a panel event study akin to difference-in-differences (DiD). We take all individuals who experience a certain event E and align their data by event time t_0 (e.g. for each person who lost a spouse, align timelines so t_0 is the widowhood start). We also include data from individuals who didn't experience that event as a comparison group.

We then estimate a model of the form: $Y_{it} = \alpha_i + \lambda_t + \sum_{k \in K} \beta_k \cdot 1[t - t_{0,i} = k] + \theta^T \text{Nuis}_{it} + \epsilon_{it}$,

where:

Y_{it} can be a variety of outcome measures: we will look at HCl itself, menu entropy E_{it} , repair stickiness S_{it} , repair, social tilt T_{it} , etc., as dependent variables. α_i are individual fixed effects (each person as their own control, accounting for stable differences in baseline happiness, etc.).

λ_t are time fixed effects (e.g., year-month dummies or something to capture any overall trends like a pandemic year affecting everyone).

The β_k coefficients capture the dynamic response at relative time k (e.g., $k = 0$ might be the event month, $k = 1$ one month after, $k = -1$ one month before, etc.). Essentially, this gives us an estimated trajectory before and after the event (with $k < 0$ as potential “pre-trends” and $k \geq 0$ as post-event effects).

Nuis_{it} are nuisance covariates like age, season, concurrent unrelated events, etc. We include tests for parallel trends: prior to the event, the treatment and control groups should have similar trends in y (no big divergence, otherwise our comparison is suspect). The coefficients β_k for $k < 0$ (pre-event) should be ≈ 0 if parallel trends hold. QS predictions for a negative shock (loss/threat): In the immediate aftermath ($k = 0$ or $k = 1$ period), we expect:

A significant drop in HCl (the event hurts, so happiness goes down sharply initially).

Then, in the following periods, a reversion toward baseline or even an overshoot above baseline as compensation kicks in (we might see β_{+2}, β_{+3} moving back up).

Menu entropy \downarrow (people focus, less variety right after a shock).

Repair stickiness \uparrow (they increasingly engage in coping or support behaviors).

Kin/care social tilt \uparrow (they might reach out more to family or friends for support).

These effects might be moderated by the severity of the shock to ledger: for instance, if the person had a large “debt” already (bad times before the event), the compensation

after might be stronger, or horizon (if someone's older or already ill, a new loss might have different dynamics than for someone with long horizon ahead).

For positive shocks: immediate HCl bump, but not unchecked: we might see smaller or shorter-lived boost than hedonic adaptation alone predicts, because QS might temper it or the person might take on new burdens (e.g. new child is joyful but also tiring). Composition might shift to prosocial or meaningful uses of the gain (higher kin/care tilt, etc.), especially if horizon is not long (the person might think: "I better use this wisely while I can").

Difference-in-differences analysis will yield an impulse response function (IRF) of the shock: a curve of β_k over time showing how the outcomes deviate from baseline around the event. We'll specifically look for that pattern of dip then recovery (for negatives) or spike then normalization (for positives), along with the composition changes.

13.3.4 Design 2: interrupted time series (ITS) within person

Another approach, especially when we have fewer instances or want individual-level insight, is interrupted time series for each person who had a sharp event. Essentially, for a given individual with a well-defined event time t_0 , we fit a time series regression:

$$HCl_t = \beta_0 + \beta_1 t + \beta_2 \cdot 1[t \geq t_0] + \beta_3(t - t_0) \cdot 1[t \geq t_0] + \gamma^T Nuis_t + \epsilon_t.$$

This includes:

A baseline trend $\beta_1 t$ (were they improving or worsening over time before the event?).

A level change β_2 at the moment of the event (jump up or down).

A slope change β_3 after the event (did the trend change trajectory?).

Nuisance controls (like day-of-week, holidays, other events overlapping).

We repeat such models for other outcomes like entropy, stickiness, etc., or do a multivariate ITS.

From ITS:

β_2 tells us the immediate impact (e.g. HCl dropped by X points at the event).

β_3 tells if there's a longer-term drift change (like before event they were stable, after event they have an upward trend indicating recovery, or vice versa).

QS evidence: After a negative shock, we might see a sharp drop ($\beta_2 < 0$) followed by a positive β_3 (an upward slope indicating improvement beyond just rebound – possibly overshooting baseline temporarily). For composition measures, β_2 for entropy might be

negative (immediate narrowing), β_3 possibly zero or positive if variety gradually returns. For stickiness or social tilt, β_2 might be positive (immediate jump in seeking support) and then maybe slope back to normal as life stabilizes. We would check if these changes are statistically significant and meaningful in size.

Each person's ITS can be combined by meta-analysis or hierarchical modeling to see overall patterns.

13.3.5 Design 3: regression discontinuity (timed policies)

Sometimes life events occur not purely by chance but via crossing a threshold – for instance, a benefit that kicks in at age 65 (retirement pension), or student loan forgiveness after 20 years of payments, etc. These situations allow a regression discontinuity design (RDD). If the threshold is arbitrary in terms of who is just below vs. just above, we can treat assignment as local random.

We set it up as: $Y_i = \tau \cdot 1[X_i \geq c] + f(X_i - c) + v_i$,

where X_i is the running variable (e.g. age or time in program), c is the cutoff (e.g. 65th birthday), $1[X_i \geq c]$ indicates the treatment (event) occurred, and $f(\cdot)$ is a smooth function modeling the underlying trend in absence of the event.

In our context, the outcome Y_i could be something like “repair stickiness” or social tilt measured over a window right after $X=c$. For example, suppose at age 65 a mandatory retirement causes people to lose a certain role (some countries have forced retirement age). We could look at well-being metrics right before and after 65 for those in jobs subject to that rule. A *positive discontinuity* in, say, time spent with family or engaging in hobbies right at 65 (compared to just younger) could indicate QS's channel-widening effect as a life chapter ends.

In general, a statistically significant jump at the threshold in a direction consistent with LoF predictions would support QS mechanisms. For instance, if at some cutoff for debt relief, we see a discontinuous increase in prosocial behavior or relief-seeking (people suddenly feel free to focus on other things), that could be evidence.

We have to be cautious: RDD requires a sharp policy or threshold that can be exploited and enough data around it. We identify any such natural cutoffs (like Medicare eligibility, etc.) that might plausibly be used. If found, and if people just above vs. just below truly differ only by the “treatment” (thanks to local randomness), it's a powerful quasi-experimental confirmation.

Example: Imagine a program that forgives medical debt for terminal patients after 5 years in hospice. If some patients hit 5 years and get relief, others at 4.9 years haven't – we

could see if hitting that 5-year mark (assuming random around threshold) causes an immediate boost in quality of life or reduction in stress signals. If LoF is real, relieving a huge burden should result in a tilt toward positive emotions or maybe a peaceful period (because the ledger got a shock improvement).

In summary, regression discontinuities offer a way to test QS predictions in scenarios where life's "interventions" are not random but based on a cutoff. A positive finding (e.g., a jump in repair stickiness at eligibility threshold) would support the idea that when a channel is opened or a burden lifted at a specific point, QS immediately exploits it (people take that opportunity to balance their ledger).

13.3.6 DAGs and confounding we must control

With observational event data, confounding is a concern. We often draw a directed acyclic graph (DAG) to think clearly:

Imagine socio-economic status (SES) influences both likelihood of an event and also directly affects well-being trajectories. For example, low SES might increase chance of job loss and also correlate with lower baseline HCI. If we don't account for SES, we might attribute differences to the event that are really due to SES.

Our DAG might look like:

SES -> Event -> { HCI, E, S_repair, T }

SES -> { HCI, E, S_repair, T }

Meds -> { ... }

This says SES impacts both the event and the outcomes; medications or other interventions might influence outcomes too but not necessarily the event.

Control strategy:

We include person fixed effects (α_i) in panel models to eliminate time-invariant confounds like baseline SES, personality, etc. That way, each person is compared to themselves.

We include time-varying controls for things like medication changes, new therapy, or anything concurrent that could bias outcomes.

For events where a comparison group is used, we can use propensity score weighting or matching: e.g. weight non-event individuals to resemble event individuals on observed covariates up to the event, ensuring apples-to-apples comparison.

In some studies (like relocation due to natural disasters), we can find instrumental variables (e.g., distance from the storm's path as an instrument for disaster exposure, assuming that near misses approximate random assignment).

The key is to block backdoor paths: e.g. controlling SES and other pre-event differences so that the observed effect is more cleanly attributable to the event. Our analyses will clearly report what covariates were controlled and run sensitivity checks for unobserved confounding (see Section 13.4.5 on R-V curves for assessing how strong an unmeasured confound would need to be to invalidate results).

13.3.7 Heterogeneity: horizon and ledger matter

Not everyone responds to events the same way. QS theory says horizon and prior ledger position modulate responses. We test interactions to capture this heterogeneity:

For example, in an event-study or ITS, we can include an interaction of post-event indicator with the person's horizon or ledger state at the time of event:
$$Y_{it} \sim \dots + \delta_1 \cdot 1[\text{Post}] \times \hat{H}_{\{t_0\}}^{-1} + \delta_2 \cdot 1[\text{Post}] \times |L_{\{t_0-\}}| + \delta_3 \cdot 1[\text{Post}] \times (\hat{H}_{\{t_0\}}^{-1} \cdot |L_{\{t_0-\}}|) + \dots$$
This looks at whether the *effect of the event* in the post period is bigger for those with short horizons or big prior imbalances.

QS predicts: If two people both lose their job, the one who is older (shorter remaining horizon) or who was already in a big hedonic deficit may show a stronger compensatory response (maybe they quickly find meaningful volunteer work or reconcile relationships, as if subconsciously knowing they need to fix things, whereas a young person might just be upset for a while but figure there's plenty of time to catch up later). So δ_1 (horizon interaction) likely positive for repair outcomes — meaning the shorter the horizon, the more post-event tilt toward repair. And δ_2 (ledger debt interaction) likely positive too — a big negative ledger pre-event might spur a larger effort to compensate after an additional hit, or could conversely flatten if the person is overwhelmed (this one is a test which way it goes). The combined interaction δ_3 could check if *both* short horizon and big debt together produce an especially pronounced response (maybe a multiplicative effect).

We will estimate these to see if, indeed, menu tilt and ledger reversion are stronger when horizons are short and pre-event debt is large. That result would align with LoF being an active constraint: it kicks into high gear precisely when it's most needed (when time is short and imbalance is high).

13.3.8 Negative vs. positive shocks: asymmetric predictions

LoF (via QS) does not claim symmetry in how negative vs. positive disturbances are handled. In fact, we predict some asymmetries:

Negative shocks (e.g. the death of a loved one, major illness diagnosis): We expect an immediate drop in HCI (sadness, stress), but *if channels are available*, a shift toward compensatory experiences afterwards. For instance, after a really bad day, that night's dreams might become more positive than usual (a counterweight) – Chapter 10 discusses dream compensation. Or in waking life, after a tragedy, people might experience moments of profound relief or meaning (not necessarily fully offsetting, but pushing upward) that aren't just regression to baseline but can overshoot it temporarily (like finding a deep sense of purpose in helping at the funeral, etc.). In data, we'd see daytime affect gradually improving and sometimes overshooting the old baseline on some days, and composition shifting (more time with supportive others, etc.). Also, dream content could show a telltale opposite valence: e.g. nightmares after trauma is common, but QS might manifest as some proportion of dreams unexpectedly benign or even positive following persistent distress, as the mind seeks relief.

Positive shocks (e.g. big win, sudden success): These produce a jump in HCI (joy, excitement), but QS would quietly ensure it doesn't result in a permanent ledger imbalance in the positive direction either. How? Possibly through *dampening mechanisms*: after an extreme high, individuals might feel a bit of anxiety or guilt (dreams might become threat-tinged or negative – a well-known phenomenon where very positive events can trigger anxious dreams). Behaviorally, a person might take on new responsibilities or worries that counteract pure elation (someone who wins money might suddenly worry about managing it or feel a responsibility to others). QS doesn't "punish" happiness, but it introduces subtle counterweights to prevent an unchecked surplus. We predict compositions shift toward care/meaning rather than continued indulgence – e.g. instead of partying endlessly after a windfall, perhaps the person soon starts focusing on family or charitable projects, especially if they have a sense of limited time. Also, dreams might show more negative or challenging themes right after huge positive events, as if keeping the ledger from running away high.

These asymmetries distinguish QS from simple hedonic adaptation. Adaptation theory would say ups or downs, people just drift back to baseline roughly symmetrically. QS/LoF suggests after extreme events, the *direction* of compensatory changes is aimed at neutralizing the net effect, which can mean overshoot above baseline for negatives and undershoot for positives (creating a balanced area under the curve, not just a return to baseline).

We will explicitly compare these predictions to adaptation: e.g. if a dataset shows that after very bad events, people not only bounce back but often experience periods of above-typical happiness (given support), that's beyond what adaptation alone claims. Or

if after very good events, there are signs of discomfort or increased striving that reduce net gain, that again is unique to a fairness constraint perspective.

13.3.9 Worked examples (what to expect)

To make it tangible, here are some example outcomes we might see if LoF holds:

Bereavement: Suppose we look at people for 90 days before and after the loss of a spouse. We might quantify: in the 90 days post-loss, their daily entropy drops by $\sim 0.3\sigma$ (they do fewer kinds of things) and repair stickiness goes up $\sim 0.4\sigma$ (if they start a comforting routine, they keep at it). Social tilt increases – they spend relatively more time with close family than they did before. HCI plummets immediately (lots of grief), but by day 45–60 we might see partial reversion toward baseline happiness levels (not full happiness, but not as low as initially). Also, perhaps their dream reports during those weeks show more “reunion” or warmth themes than usual – a possible counterweight to daytime grief.

Windfall: For someone who wins a large lottery or gets a big sudden financial gain, initial HCI spikes (elation). We predict that doesn’t last unmitigated: maybe within weeks, their adversarial share does *not* rise (they don’t become more hostile or selfish – in fact they might become more prosocial), and they might increase prosocial outflows (gifts, helping others). If that person is older, they might turn to legacy tasks (like estate planning, philanthropic actions) pretty quickly, making those “sticky” behaviors.

Cancer diagnosis: An individual gets a serious diagnosis. Immediate effect: HCI falls (fear, sadness). But then a critical factor is channel access: if they have good pain management, support, etc., we’d expect some rebound in mood over the following weeks – perhaps not to prior baseline, but an adjustment. The channel index (sleep, pain control, social support available) might predict *how fast* they rebound. If channels are open, QS can work; if not, they might stay down. In a research setting, we might even measure brain signals: their next annual fMRI could show that vmPFC (valuation area) responds more strongly to “repair” options than it did before, indicating a heightened drive to seek positive balance after the shock.

These are hypothetical numeric examples, but they illustrate the type of evidence we’ll gather: effect sizes of changes in behavior and affect after events, and qualitative shifts in things like dream content.

13.3.10 Minimal preregistration package

To ensure credibility, we preregister a plan for analyzing life events, including:

Event definitions and timestamps: Exactly how we define each event type and where t_0 comes from (e.g. “bereavement = date of death of someone in circle of closeness 1 or 2” and how that’s recorded).

Primary outcomes: We list what we’ll measure as responses. For example: HCI (hedonic level), entropy E, repair stickiness S_repair, and social tilt T as our four main ones. Possibly also something like nightly dream valence or channel index if relevant.

Windows: The time window for analysis – e.g. we’ll examine [−90 days, +180 days] around each event, meaning 3 months prior to 6 months after, or adjust as suitable per event (some events might have longer-term effects).

Models: Which analysis methods we’ll use – event-study DiD and/or ITS, as described.

SESOI (smallest effect sizes of interest): We set target effect size ranges: say we consider changes of $d = 0.15\text{--}0.25$ (15–25% of a SD) in composition metrics as meaningful, and maybe $d = 0.20$ for volatility shifts. If observed effects are smaller than these consistently, we might call it inconclusive or no strong support.

Negative controls: We plan checks like using anticipated-but-never-happened events (someone thought they might lose their job but didn’t) or “placebo dates” randomly chosen that should have no effect. These should show no pattern; if they do, something’s off.

Subgroup analyses: We’ll look at subgroups by horizon (e.g., younger vs. older individuals experiencing the same event) or ledger quartiles (someone who was already very unhappy vs. very happy before the shock).

This preregistration ensures we’re not data-dredging. We specify ahead how we’ll judge if LoF’s predictions pan out.

13.3.11 Ethics and dignity

Studying life events needs ethical care. While we’re mostly observing natural occurrences, we still:

Do not stage events. We obviously do nothing to cause or encourage events (we’re not manipulating lives, just observing what happens). We also do not push people to talk about or log things if they are in a sensitive moment; we let them control data sharing especially during crises.

Provide standard supports: If someone goes through a crisis, our role as researchers doesn’t trump basic humanity – participants receive or are pointed to normal help (e.g. if our system detects extremely high distress or a flagged event like bereavement, it can

gently offer resources such as counseling contacts or simply reduce the burden of surveys). As mentioned, we might have features like sleep or pain prompts and communication tools (like an easy way to notify their support network) that are available to everyone irrespective of being in an “event” or not.

Privacy around events: Data during sensitive periods are handled with extra care. Perhaps participants can mark an entry “don’t use this detail” or can choose to mute certain streams when something very personal happens – we respect that, and we model the missingness (see Section 13.4) rather than intrude.

No exploitation of vulnerability: We explicitly commit that if someone’s going through a hard time, we don’t accelerate data collection or probing questions just because “this is interesting data.” The dignity and autonomy of participants come first. If anything, the app likely *reduces* requests during those times (unless the participant wants to engage as a form of coping or diary).

Feature-only data: If someone’s life event involves others (e.g. a conversation with a family about a death), we ensure we are not inadvertently capturing identifiable info about third parties without consent. We only log broad features (“big event today” toggle, or counts of calls) and never content.

The guiding principle is we observe and support, never interfere or add burden during life events. This way, any patterns we detect come from the person’s natural response, not from our involvement.

13.3.12 What would count against QS

Finally, we outline what outcomes in the event analyses would constitute a failure of QS/LoF predictions:

Suffering without compensation: If large negative shocks lead to persistent increases in adversarial or avoidant behavior, and no rise in repair activities, even when support channels are available, that’s a direct contradiction. For instance, if after a major loss, some individuals just spiral into long-term bitterness or disengagement with no compensatory positive period, that is problematic for LoF.

Unmitigated indulgence: If large positive shocks result in some people engaging in prolonged, uncompensated indulgence with no counter-balancing adjustments (and no eventual return toward baseline or concern for others), that would violate the expected dynamic. It would suggest the system allows persistent surplus in some cases.

No dream counterweight: If we have data on dreams and we find that after tough days or events, the dream affect shows no inversion (e.g. it’s just as negative as the day, rather

than lighter), that would remove one hypothesized compensation channel (as detailed in Chapter 10).

Rival explanation fits as well: If standard psychological models (hedonic adaptation, or say increased risk-aversion after bad events) can fully explain the trajectories without invoking horizon or QS, and those models predict the data comparably or better, then QS isn't pulling unique weight. For example, if an adaptation model with a set-point explains all rebound effects and adding a horizon term doesn't improve anything, then the events haven't revealed a distinct LoF signature.

If patterns like the above were *replicated across multiple event types and people*, it would force us to retreat from calling LoF a law, perhaps viewing it just as one of many tendencies, or rejecting it if completely inconsistent.

Takeaway: Life already provides numerous “experiments” in the form of dramatic events. By analyzing these carefully (with proper controls and models), we can test whether people’s responses align with a fairness constraint. Do their actions and experiences shift in a way that tends to restore balance? This section sets up how to find out. If we see, time and again, that after life knocks you down the system quietly lifts you up (and vice versa), that’s strong evidence for LoF. If instead we see some people just accumulating net suffering or coasting on net joy with no correction, that challenges the universality of the law.

13.3.13 Where we go next:

Section 13.4 tackles the *statistical backbone* needed to make these conclusions credible: dealing with missing data and using hierarchical models to separate real signals from noise without fooling ourselves.

13.4 Research Notes: Missingness and Hierarchical Models

Real life data are messy. Phones die, wearables get left on the nightstand, diaries go unfilled, people drop out for weeks, and—importantly—people choose privacy over data sometimes (as they should). If we pretended those gaps don't matter or just did a simplistic analysis on whatever data we happen to get, any claim about fairness would stand on sand. This section lays out a rigorous approach to handle missingness and to model individuals and groups with hierarchical structures so that (i) our estimates remain honest (not biased by who reported when), (ii) uncertainty from data holes is properly propagated (so our confidence intervals reflect what we *don't* know), and (iii) conclusions don't hinge on rosy assumptions (we stress-test them against worst-case scenarios).

13.4.1 Name the missingness (before you analyze)

The first rule of handling missing data: know why it's missing (or at least make your best guess). We classify missingness mechanisms using standard terminology (MCAR, MAR, MNAR) and tie each to real examples, *before* running analyses:

MCAR (Missing Completely at Random): The chance of data being missing is unrelated to anything (observed or unobserved). Example: a sensor vendor's cloud service has a random outage one day, dropping everyone's data for that day. Or a server error that randomly fails to record some entries. This is rare in practice, but if it happens (and we verify it's truly random), it's the easiest to handle (just ignore those points or impute with any unbiased method).

MAR (Missing At Random): Missingness depends only on *observed* data. Example: We find that older participants tend not to carry their phones after 8pm, so evening EMA are missing more for older folks; or on days when someone goes to the clinic (which we know from schedule or health records), their step count is lower or missing because they left their wearable at home. These can be addressed by conditioning on those observed factors (age, or an indicator of clinic-day) in the model. As long as we include those, the missingness can be considered ignorable in estimation.

MNAR (Missing Not At Random), a.k.a. Informative Missingness: The probability of missing data depends on something *not observed* – often the thing we wish we knew. Example: People skip the evening mood survey specifically when they are very distressed (we observe no data precisely when their HCI would have been low). Or participants do not report dreams on nights they have nightmares, because they prefer not to record them. This kind of missingness is the default in affect research (people often disengage when feeling bad), and we cannot just ignore it. It must be explicitly modeled.

Our rule: *Unless* we have strong evidence to justify MCAR/MAR (using logs, self-reports, or patterns that indicate missingness is unrelated or only related to known data), we will assume informative missingness (MNAR) for critical streams like EMA, dream reports, communications frequency, etc. and use models that account for it. In simpler terms, we give missing data the benefit of the doubt: it probably means something, often that the person's state influenced their reporting, so we can't treat it as random noise.

We document in the preregistration which streams we suspect are MNAR and why, and we outline how we'll model those missingness mechanisms. For example, "Evening EMA: likely MNAR because participants say they 'don't feel like answering' when sad. We will model response probability as a function of latent daily affect."

By naming it upfront, we avoid post-hoc rationalizations and ensure the analysis plan includes the right adjustments.

13.4.2 Joint models that respect informative gaps

The gold-standard way to handle MNAR is joint modeling of the data process and the missing-data process together. That way, the model itself corrects for any bias introduced by missingness.

We implement a shared-parameter joint model: one part for the outcome of interest (e.g., daily HCl) and one part for the missingness (whether HCl was observed).

Outcome model (state-space):

This is essentially the longitudinal model we already described for HCl. For example, for person i :

$$x_t^{(i)} = \phi x_{t-1}^{(i)} + b^T Z_t^{(i)} + u_i + \eta_t^{(i)},$$

$$HCl_t^{(i)} \sim \mathcal{N}(x_t^{(i)}, \sigma^2_{\text{meas}}),$$

as earlier. This formalizes that latent affect x_t follows some process (with ϕ autoregression, covariates Z such as weekdays or intervention indicators, person random effect u_i , etc.), and observed HCl is a noisy measurement around it.

Missingness (selection) model: We model $M_t^{(i)} = 1$ if HCl is missing at time t for person i . A logistic regression could be:

$$\Pr(M_t^{(i)} = 1 | x_t^{(i)}, W_t^{(i)}) = \text{logit}^{-1}(a + \gamma x_t^{(i)} + c^T W_t^{(i)}),$$

where W_t are any observed predictors of missingness (such as battery level or whether it's a weekend). The key term is $\gamma x_t^{(i)}$ — this captures informative missingness: if $\gamma \neq 0$, it means the person's latent affect directly influences the likelihood of missing data. For

instance, if γ is negative, when x_t (affect) is low (bad mood), the probability of missingness M_t increases (the logit is more likely). This matches the scenario where individuals skip reports when feeling down.

In this joint model, the outcome and missingness share the latent x_t . We fit it all together, typically in a Bayesian framework (because it's complex and we want full posteriors). The payoff: if indeed people skip when feeling bad, the model will infer a negative γ , and effectively it will "correct" estimates of average mood by realizing the observed data overrepresents good moods (since bad mood days were underreported). It will then down-weight the overly rosy data or adjust latent x_t distributions to fill in the gaps.

We interpret γ : if we find γ significantly nonzero (especially if negative), we confirm MNAR. If $\gamma \sim 0$, our data suggests missingness wasn't strongly dependent on affect after controlling other factors, which is interesting itself.

We also consider an alternative approach if needed: pattern-mixture models. This means rather than explicitly linking to x_t , we stratify trajectories by missingness patterns (e.g. people who missed >50% vs. <50% or specific sequences) and allow each pattern to have its own outcome distribution. We can then add "delta adjustments" to simulate worse-case values for missing segments (described in 13.4.6). Pattern mixture is another way to be robust: instead of modeling the cause of missingness, it just acknowledges different "types" of respondents and sees how conclusions vary.

We will fit joint models in a Bayesian way so that uncertainty flows through. This means when missingness is high, the posterior for the ledger or effect sizes will widen to reflect that we're less sure. All that uncertainty—due to data we didn't get—*must* carry forward into final tests of the law, otherwise we'd be overstating what we know.

13.4.3 Hierarchical partial pooling: why it matters

Our dataset spans multiple levels: repeated measures (days) nested in persons, persons possibly nested in sites or cultures. Hierarchical modeling (multi-level modeling) explicitly accounts for this structure, which improves inference by "sharing strength" appropriately and avoiding both overfitting and misestimating variability.

We include random effects at relevant levels:

Level 1 (day-to-day): the day-level model we already have for $x_t^{(i)}$.

Level 2 (person): each person i can have random intercepts (we had u_i above for baseline affect) and possibly random slopes for key effects. For example, one person might consistently react more strongly to horizon changes than another. We might say a person's coefficients β_i for certain predictors are drawn from a population distribution

$\mathcal{N}(\mu\beta, \Sigma_\beta)$. Similarly, $u_i \sim \mathcal{N}(0, \tau_u^2)$. Level 3 (site/culture): If our study spans different locations or cultural contexts, we can introduce a random effect s_j for site j (with $s_j \sim \mathcal{N}(0, \tau_s^2)$). This might capture that, say, overall happiness levels differ by country or some aspects of QS might vary by culture.

Partial pooling via these random effects reduces overfitting: for example, if one person has sparse data, the model will borrow info from the group to avoid extreme estimates for that person. It also “allows for idiosyncrasies” – not forcing everyone to have identical responses, which is important because LoF must hold *within* each stream but can still vary in how it manifests.

We also allow cross-level interactions: For instance, let the effect of horizon on variance itself vary by person and site: $\beta_H^{(i)} = \mu_H + r_i + s_{j[i]}$. So person i in site j has their horizon effect = average μ_H plus a personal deviation r_i plus a site deviation $s_{j[i]}$. This means we’ll check if horizon-sensitivity is consistent—if LoF is robust, we expect to see those random slopes for horizon still mostly negative for variance across individuals (maybe varying in magnitude, but rarely positive). If horizon effect varied wildly (positive for some, negative for others), the “law” would be more like an “option.”

For compositional outcomes (like how one’s time is divided among activities), we might use a multinomial logistic model with person-level random effects tied to the latent state. We mention “correlated random effects tied to x_t ” meaning we model, say, that on days when x_t is certain value, the probabilities of being in repair vs. neutral vs. adversarial mode depend on some person-specific tendency, but we ensure that structure still borrows across people. This prevents, e.g., claiming a “tilt” just because one person had unique habits—if only one person in our sample always goes adversarial after events and no one else does, partial pooling will treat that as an outlier rather than evidence against LoF (or it will widen uncertainty accordingly).

13.4.4 Imputation that respects dynamics

When data is missing, one approach is imputation – filling in or simulating the missing values to complete the dataset and then analyzing as usual. But a naive imputation (like “carry last observation forward” or average of neighbors) can distort dynamics, especially if data isn’t missing at random. We avoid any simplistic patching. Instead, we generate imputations from the posterior predictive distribution of our state-space model, which inherently respects temporal dynamics and uncertainty.

Concretely:

For continuous sensor streams (e.g. heart rate, sleep minutes) with gaps, we can use Kalman smoothing or Gaussian Process interpolation. These approaches leverage the correlations in time to predict likely values and give an uncertainty band. They won't just flat-fill the mean; they give a distribution of possible trajectories that align with observed data points.

For count data like number of messages or calls, we can use inhomogeneous Poisson processes or Hawkes processes to simulate plausible event times in gaps. For example, if we know someone usually texts 5 times/day and often in bursts, a Hawkes model can simulate realistic patterns in a missing interval.

For ordinal or Likert EMA outcomes, we use ordinal regression models to draw imputed values, possibly with a probability of zero if "silence" is common (some days they just don't report – a kind of zero inflation). We might incorporate knowledge like: person tends to skip if feeling bad (the joint model handles that and will produce more low imputed values on missing days accordingly).

Cross-modal borrowing: If one stream is missing but others are present, our model can lean on them. E.g., no self-report tonight but we have high EDA (stress indicator) and very few steps – the model might infer x_t was low and impute correspondingly. The uncertainty will be larger than if self-report was there, but narrower than if *all* channels were missing. Essentially, the joint model's latent state serves to propagate info from observed modalities to unobserved ones.

We don't do just one imputation – we do multiple imputation via posterior draws. We will generate, say, 10 or 20 completed datasets by drawing from the joint posterior of the model, which gives us a distribution of possible values for each missing entry. We then run our analyses (like testing compression or event effects) on each and combine results (e.g., Rubin's rules for multiple imputation, or in a Bayesian workflow, pooling posterior draws across imputations). The end result: our effect estimates already incorporate uncertainty from missing data. For instance, if there was a chance that all those skipped pain reports could mean the person was actually in a lot of pain, our final ledger confidence interval will reflect that (it'll be wider toward the negative side).

Furthermore, any reported ledger total $L(T)$ will come with an uncertainty that includes missingness uncertainty. If someone only reported mood half the time, the error bars on their final ledger will be big. That honesty is crucial: one might "find neutrality" only to realize the CI is huge due to missing data, meaning we can't really claim it.

13.4.5 Diagnostics that catch wishful thinking

We put in place several diagnostics to ensure our models aren't tricking us into seeing a pattern that isn't real:

Missingness regressions: We will visually and statistically check how missingness relates to observed data around it. For instance, plot the probability of missing tomorrow given today's mood, or missingness vs. next day's mood. If we see missingness spikes *before* adverse shifts (e.g. people stop reporting right before a depressive episode), that strongly indicates MNAR and justifies the joint model. We also check if our missingness model fits well (e.g., does it predict the observed pattern of skips).

R-V curves (Robustness–Value curves): These come from the causal inference literature (sensitivity analysis for unmeasured confounding). They tell us: how strong would an unmeasured confounder have to be to wipe out the effect we see? For missingness, we might ask: how bad would hidden bias have to be to explain away the horizon-compression effect? We vary an assumed γ or missingness mechanism strength to see at what point the effect goes away. If it would require something implausible (e.g., "people only report on their very best days and never on average days"), then our result is robust.

Posterior predictive checks: We simulate full datasets from our model (including generating fake missingness and fake observed values) and compare to actual data. Are the properties similar? For example, do simulated time series have similar volatility, bursts, weekend vs. weekday differences, distribution of dream reports, etc.? If our model can't reproduce basic features of the data, we mistrust its specific predictions. We tune or choose a better model.

Leverage and influence diagnostics: Check if our results (like the magnitude of horizon effect) are being driven by a small subset of participants. We can do a leave-one-participant-out (or leave-one-site-out) analysis: re-estimate key effects excluding each person (or site) one at a time to see if any omission drastically changes the estimate. If, say, removing Participant #42 who contributed a ton of data causes the compression effect to disappear, then our evidence is too dependent on that one person – maybe they were special. In that case, we'd report that and be cautious. Ideally, effects hold even when any single cluster is removed.

In short, we actively look for *signs that we might be fooling ourselves*. If the horizon effect only appears because one heavy user provided most of the data, or if our model strangely predicts patterns in missingness that don't hold, we need to know and adjust.

13.4.6 Sensitivity analyses you can read in plain English

We will conduct and publish sensitivity analyses in a form that non-statisticians (and our readers) can understand, to demonstrate how robust the results are:

Delta-adjustments (pattern-mixture approach): We assume in turn that all missing entries in certain key streams were some δ amount worse (or better) than observed ones and re-run the analysis. For example, we might assume that every missing mood report is 0.2 SD lower than it would have been if observed (a moderately pessimistic scenario). Then see if our main conclusions (like “variance compresses at end-of-life” or “post-bereavement stickiness rises”) still hold. If LoF claims only hold when we assume missing data were neutral or positive, but collapse when we assume missing data were more negative, that would be worrying. We likely will test a range: $\delta \in \{0.1, 0.2, 0.3\}$ SD worse, and maybe symmetric for better (though usually missing when bad is the concern). We expect LoF signatures to survive modest deltas – meaning even if missing times were somewhat worse, the effect is still there. If it requires wild deltas to break it (like assuming every skip was an extremely terrible day), we’ll report that threshold (e.g. “the horizon effect disappears only if unreported days were on average >0.5 SD worse than reported days, which seems unlikely”).

Tipping-point analysis: Specifically for the selection model coupling (γ in 13.4.2): we increase $|\gamma|$ to see how strong the missingness-affect relationship would need to be to nullify a given result. For example, “How strong would the bias have to be (people only reporting on good days) such that the end-of-life variance drop could be explained away as an artifact?” We would report something like: “if people skipped reports whenever their HCl was below the 20th percentile, then the observed compression could be fake; but our analysis of partial data suggests missingness wasn’t that extreme.” Or more formally: we find the γ at which the 95% CI of the horizon effect crosses zero, and say “Our results are robust unless non-reporting on bad days is very highly pronounced (e.g., participants would have to not report on 80% of moderately bad days).”

By publishing these in straightforward terms, we make it clear where the weak points could be. If an effect only holds by assuming something convenient, that will be transparent.

We also include negative-control outcomes as another kind of sensitivity: e.g. if we run the whole pipeline but instead of HCl outcomes we use something irrelevant like weather in another city, we should get null results. If our pipeline “finds” patterns in random or unrelated data, then it might be overfitting or leaking, which invalidates the real findings. We’ll report that we tested these pipeline sanity checks.

13.4.7 Priors and regularization that prevent overclaiming

In Bayesian analyses, priors can help stabilize estimates, especially in complex models. We choose priors that are *weakly informative* and bias us away from extreme or too-good-to-be-true results unless the data really support them:

We put regularizing priors on variance parameters: e.g. for random effect standard deviations τ , a Half-t with small degrees of freedom, like $\tau \sim \text{Half-t}(3, 0, 0.5)$. This slightly pulls down implausibly large heterogeneity unless data demand it. So if one person appears to have an astronomically different pattern, the model is slightly skeptical unless evidence is strong, which prevents over-interpreting noise as huge person-specific effects.

Sparsity priors (like the horseshoe) for large predictor sets: If we include many candidate features in telemetry, we encourage most to have near-zero effect unless data strongly indicate otherwise. This helps because with rich data it's easy to find many tiny "significant" effects – a horseshoe prior on, say, dozens of behavioral features ensures we only highlight those that truly stand out.

Hierarchical priors on missingness coupling (γ): We are cautious about the MNAR correction not "over-explaining" things. If the model has a lot of freedom, it could, in theory, attribute everything to missingness (like fit the data by saying "whenever we didn't see you, you must have been super miserable, thus no fairness effect needed"). To guard against that, we might put a mild shrinkage prior on γ around 0 (e.g., $\mathcal{N}(0, 1)$ or something) hierarchically if multiple people are involved. Essentially, we don't let the model assume crazy missingness effects unless the data insist. This protects against the model "explaining away" a real signal with an extreme, speculative missingness relationship.

All priors will be published along with prior predictive checks – we simulate from the priors to show they are not encoding anything too strong (like we're not a priori forcing variance to always drop or anything). We calibrate them so that before seeing data, they allow a wide range of plausible outcomes, just ruling out ones that are nonsensical (e.g., an effect size of 100 SD or negative variances, etc.).

In summary, these priors function like guard rails: they keep our model from going off the rails on the basis of limited data or chasing patterns that might just be noise. They ensure that any final claims are earned by the data.

13.4.8 Putting it together: the analysis pipeline

Bringing all these pieces together, our analysis pipeline for the longitudinal data is:

Pre-analysis audit: We start by summarizing what data we have – how complete each stream is, obvious outages (e.g., “fitbit server down on Jan 3-4 for all users”), and any patterns like “20% of participants never did any dream logs” etc. We label each stream with what we suspect (MCAR, MAR, or MNAR) as per 13.4.1 and note any specific mechanism (e.g., “evening reports likely MNAR due to fatigue”).

DAG and nuisance plan: We lay out which covariates will be included to block confounds (e.g., we’ll include age, baseline health, a random intercept per person, etc.), which variables might serve as instruments if needed, and which will be used only in sensitivity analysis but not in primary models.

Fit hierarchical state-space joint model: Using something like Hamiltonian Monte Carlo, we fit the model described in 13.4.2–13.4.3. This model simultaneously estimates latent HCI paths and the missingness parameters.

Generate multiple imputations: We draw, say, $M=100$ samples from the posterior of that model (or fewer if it’s heavy) and for each, we simulate a complete dataset (fill in missing values with draws from the model). Now we have many plausible completed datasets.

Compute derived measures: For each imputed dataset, we compute daily variance, entropy, stickiness, ledger $L(t)$, etc., and then aggregate to the level needed for tests (maybe per person end-of-life metrics, etc.).

Conduct main tests with uncertainty: For example, test “is variance lower in last month vs. before” by looking at the distribution over imputations and individuals. Or fit a simpler regression on each imputed set (like horizon vs. variance) and combine results. Essentially, get a posterior or sampling distribution for each key hypothesis (horizon effect, event effect, etc.) that incorporates missingness uncertainty.

Run diagnostics and sensitivity: We take the fitted model and data and do all the checks from 13.4.5 and 13.4.6. Plot missingness vs. observables, do leave-one-out checks, generate R-V curves, try delta adjustments (which might be as easy as tweaking γ in the model and refitting or just adjusting outcomes in completed data to mimic an extreme).

Produce visualizations: e.g. a chart showing predicted vs. actual variance over time, or how robust the result is across the δ range, etc.

Lock everything and share: We *pre-registered* this pipeline, but now we also freeze the code (perhaps make a git commit or hash it) so we know exactly what produced the results. We prepare a preprint or report documenting the pipeline. Crucially, we also generate a synthetic dataset (no real user data, but simulated from our model) that has similar statistical properties. This synthetic dataset, along with the code, can be released

so that others can run the entire analysis end-to-end and verify or play with it without privacy concerns.

This might seem heavy, but it's what it takes to make claims about something as profound as a law of experience. Every step ensures that if we see fairness patterns, it's because they're truly there, not because we massaged data or ignored inconvenient bits.

13.4.9 Privacy by design (so participation isn't a filter)

A subtle issue: we don't want our study to become so invasive or creepy that only a certain kind of person participates (or people alter behavior due to being observed). That itself could bias results (if only extremely open quantified-self enthusiasts join, that's not everyone). We therefore double-down on privacy by design:

Feature-only capture: As introduced earlier, we transform raw data into summarized features *locally*. For example, instead of uploading full chat logs or audio, the app might count “you had 4 calls > 60 seconds with contacts marked ‘family’ today” and upload just that count. Raw GPS might become “radius of gyration = 5 km today”. This way, sensitive content (what exactly was said, where exactly you went) stays on the device.

Local differential privacy: For some features we even add noise on the device. For instance, if we are counting messages, we might add a random small noise with known distribution (so one day it might report 5 instead of 4 messages). Over many days, the noise cancels out for analysis, but an attacker couldn't be sure exactly how many messages on a given day. We track a “privacy budget” for each participant to make sure we're not accidentally revealing too much with repeated queries.

Federated learning where possible: Instead of centralizing raw data, we could send model code to the phone, have it update a person's parameters using their data, and then send back only the parameter updates or summary statistics (aggregated in a way that the server never sees individual data). This is cutting-edge, but we can attempt for simpler parts like calibrating a classifier on-device.

Data trust governance: We may involve a neutral third party or a “data trust” which holds the de-identification keys (so the research team never sees actual identities) and approves any data releases. The data trust could include participant representatives. Essentially, identities are separated from data by an independent steward.

Kill switch and transparency: Participants have a literal “Delete my data” button which, when hit, triggers deletion of their data from our servers (to the extent possible) and sends a log to an auditable registry (so it can be confirmed we did it). The system is designed to comply promptly and thoroughly.

The point is twofold: ethically it's the right thing to do, and methodologically it's important because it reduces self-selection bias. If people trust the system, a wider variety will participate, including those who deeply care about privacy. That means our sample is more representative. Also, if people are less worried about being watched, they might behave more naturally, giving more valid data.

By making privacy protective measures integral, we hope to avoid the scenario where missingness correlates with “did something sensitive and doesn’t trust the researchers.” That’s a tough MNAR to fix after the fact – better to prevent it by design.

We believe this dual focus on dignity and data integrity actually aligns: when participants feel comfortable and respected, we get better science.

13.4.10 What would invalidate claims despite fancy models

We can do everything right statistically and still be fooled if reality doesn’t cooperate. We list here some *deal-breakers* that, if they occur, mean we should not claim evidence for LoF:

Incoherent posteriors across model variations: If different plausible models (selection model vs. pattern-mixture vs. delta adjustments) lead to wildly different conclusions about the key effects, then we have not truly pinned down a result. For instance, if one missing-data approach says “yes, variance compresses significantly” and another approach (also justifiable) says “no, it doesn’t at all,” then we can’t, in good faith, claim compression is evidenced. In that case, the result is too model-dependent to be reliable. We’d have to report that we don’t have a consistent answer.

Permutation tests show patterns even when they shouldn’t: For example, if we randomly permute horizon labels among participants or shuffle event times, and our analysis pipeline still finds “compression” or “tilt” effects of similar magnitude, then what we found is likely a spurious artifact (maybe due to seasonality or some unaccounted trend). We’d then know our analysis is picking up something other than the theoretical effect. If such tests (Model-X knockoffs for features, random horizon assignments, etc.) reveal that patterns could be obtained under null conditions, we must not claim those patterns are proof of LoF.

External replications fail: If other teams (using our shared code and synthetic data, or running their own similar studies) cannot reproduce the key findings—despite following the protocol and having decent power—then our findings might be sample-specific or due to hidden biases. If multiple external attempts in different contexts (maybe a similar study in another country or using different sensors) come up empty or opposite, we have

to question the universality of LoF. One replication failing might be just differences; but a pattern of failures would be a strong signal.

In any of these cases, we'd have to retract or downgrade our claims. That could mean saying "We thought we saw compression, but it appears to have been an artifact of X, so at most we can call it a tentative tendency" or outright "there's no evidence for LoF in these data after all." We commit to doing this publicly if needed. The credibility of the theory demands being willing to let it go if robust evidence won't support it.

Takeaway: The Law of Fairness rises or falls on whether our analyses honor the holes and noise in the data and still find a signal. By treating missingness itself as information, using hierarchical models to borrow strength without overconfidence, and publishing exactly how fragile or robust our conclusions are, we ensure that we're not seeing "fairness" just because we looked the other way when data went missing. In practical terms, this means we won't claim LoF is confirmed just because the participants who kept their phones charged looked balanced; we will only claim it if the entire evidence, gaps included, supports it.

13.4.11 Where we go next:

Next, Section 13.5 shifts focus outward: how do we involve the public in gathering long-horizon data without creeping them out or violating trust, and still get scientifically useful results?

13.5 Citizen Science Without the Creepiness

The Law of Fairness needs decades-long, real-life data. That's more than any small research team can gather alone. The idea of citizen science is appealing: invite people everywhere to track their well-being and contribute to a massive dataset. But this immediately raises red flags about privacy, surveillance, and exploitation. How do we harness widespread participation ethically, so it's voluntary, transparent, privacy-preserving, scientifically useful, and even beneficial to participants? This section outlines a blueprint for exactly that—a citizen science approach that people can actually live with, which yields high-quality data without turning life into a creepy lab.

13.5.1 Principles people can actually live with

Any public-facing project must follow principles that respect participants as partners, not subjects. Ours include:

Consent you can see: We present consent information in plain language with a “nutrition label” style breakdown for each data type: what we collect, why we need it, how long we keep it, who can see it. No buried surprises in a 20-page TOS. At any time, a participant can review these terms in-app and see exactly what categories of data they’re sharing. This transparency builds trust.

Local first: As echoed from privacy-by-design, raw streams (like audio, GPS, full text) stay on the device by default. Only derived features (counts, averages, flags) leave the device. For example, the app might analyze your messages locally to count how many were to family versus others, and only that count is sent. This greatly reduces the risk and the feeling of being “watched”—because in a sense, you aren’t; the raw you stays with you.

Minimalism: We only collect the minimum needed to test LoF’s predictions. No always-on microphone, no random camera snaps, no detailed browser history. If it’s not directly relevant to affect or compensatory behavior, we leave it out. This not only protects privacy but also reduces burden and potential for misuse. For instance, geolocation: we decided precise GPS trails aren’t necessary, so we may only log something like general radius of movement or number of distinct places visited—enough for a “variety of environment” metric, but not enough to reconstruct someone’s exact daily route.

Revocable at will: Participation is not a trap. If someone wants to pause or stop, it’s one tap. And crucially, doing so does *not* break any core functionality of the app or penalize them. If our app doubles as a personal mood tracker or diary (to give value to users), that core function should still work in “pause” mode (just not sending data to us). No guilt trips, no nagging – if you need a break, take it. This ensures people don’t feel they’ve sold their soul to science with no way out.

Dignity by design: The app interface is designed under the assumption that participants are intelligent partners, not data sources or patients to be managed. That means, for example, using respectful language (no overly cheerful “Hi subject #123, how are we feeling?” but maybe a more neutral or empowering tone). Feedback given to participants (more on that later) should be affirming and optional – it’s their data about their life, we’re just crunching it. The experience should make them feel *curious* or *insightful* about themselves, not judged or prodded.

Open methods, closed identities: We pledge to make all our methods (the code, the models, the analyses) open source and publicly accessible, so anyone can scrutinize how we’re using the data. But at the same time, personal identities and any raw content remain strictly confidential and not public. So, it’s transparent *what* we do, but absolutely private *whom* we do it with. Participants can verify algorithms and results without seeing each other’s personal details.

These principles are communicated to the community to set expectations. They also act as internal design guidelines. If a proposed feature violates one, we don’t implement it.

13.5.2 What we would collect (and what we won’t)

Staying minimal and relevant, here’s the scope of data for citizen scientists:

We collect features, not content:

Sleep: total sleep duration per day; perhaps a sleep variability score (how consistent bedtime is) and fragmentation index (how restless/interrupted the sleep was). Captured via phone motion or wearables but summarized as numbers.

Activity: simple measures like step count, minutes spent sedentary vs. active, maybe HRV summary if wearing a sensor, but again as daily totals or stats.

Communication patterns: number of calls or messages and optionally their lengths, categorized by relationship (the user can tag certain contacts as “family,” “friend,” etc.). No recording of conversations or reading texts – just counts by category. E.g., “5 messages sent to Family, 2 to Work, 0 to others today.”

Self-reports: brief mood check-ins (valence and energy level, as mentioned in Section 13.1) perhaps once or twice a day, which take ~30 seconds. Also a nightly prompt to tag your dominant dream if you remember it: just choose if it was mostly pleasant, neutral, or unpleasant and maybe a couple theme tags (we provide a list like “chased,” “flying,” “loss,” “reunion” and the user can tick if relevant). No dream journals need be shared, just these tags.

Context toggles: Simple yes/no inputs for context that only the participant would know: “Big event today?”, “Pain above normal today?”, “Caregiving day?”. These are voluntary but give data context – e.g., someone can flag that something significant happened (good or bad) without detailing it, just so we know there was a perturbation.

We do not collect:

Precise GPS trails or continuous location tracking (too invasive, we stick to coarse metrics).

Microphone audio or ambient sound recordings (can capture sensitive conversations).

Camera feeds or facial expressions (not needed and very private).

Message content or actual text conversations, or any logging of what was said on calls.

Browser or app usage logs.

Financial transactions or records (unless a participant manually exports a summary for their own use – we don’t hook into bank accounts).

And importantly, nothing about third parties without consent – for example, we won’t ask a participant to report on their spouse’s feelings or anything like that, and we avoid features like scraping their social media which involves others’ data.

By explicitly stating what we won’t do, we set a boundary that participants can be comfortable with. It also draws a line for us: those off-limits data might be tempting for some analysis, but we choose principles over maximizing data.

13.5.3 Three participation tiers (opt-in, modular)

Not everyone will want to engage at the same level. We define tiers so people can choose their comfort zone:

Tier A: “Light” – This might be just 1–2 mood check-ins per day, plus passive basic sensors like steps and sleep from phone (if they carry it). No extra devices, no complex tasks. Dream tagging only if they feel like it. It’s basically like using a simple mood tracker app that also counts steps.

Tier B: “Standard” – Includes Tier A plus a bit more: e.g. user-curated communication tracking (they tag a few contacts as family/friends so we can count those interactions), a weekly short task like a 3-minute game or “horizon quiz” to gauge time perspective, and maybe optionally connecting a wearable for HRV. It’s more data, but still not too burdensome.

Tier C: “Methods helper” – These are enthusiasts who opt in to everything Tier B offers and then some lab-like studies a few times a year (for instance, they might do an at-home EEG session or go to a partner lab for an fMRI once a year). They might also get a stipend for the bigger effort. Essentially, these participants help us with the hardest tests (like end-of-life monitoring or experimental tasks).

The tiers are modular and interchangeable: participants can move *up or down* tiers any time. If someone starts gung-ho at Tier C but later feels it’s too much, they can drop to B or A without quitting entirely. Or if someone at Tier A finds it interesting, they might upgrade to Tier B to get more feedback or contribute more.

This tiered approach respects personal boundaries and life changes. It’s not one-size-fits-all. Also, by having a large Tier A base, we get breadth (lots of people, minimal data each) and by having some Tier C, we get depth (rich data on fewer folks). Both help.

13.5.4 Privacy architecture (how we keep our promises)

We’ve promised privacy; here’s how we technically implement it:

On-device processing: As mentioned, the app handles raw data. For example, it might listen to accelerometer all day to detect sleep vs. wake, but it only sends us “slept 7.2 hours, 2 interruptions” at day’s end. The code for those algorithms can be open source so people trust it’s not secretly uploading more.

Differential privacy at the edge: We calibrate noise for certain metrics. For example, the app might add Laplace noise to the count of messages before uploading, achieving a certain ϵ privacy guarantee that an extra message won’t significantly change reported data. The privacy budget (how much noise vs. utility trade-off) can even be shown to the user. If a user is extremely privacy-conscious, they might choose a setting that adds more noise (with a note that it slightly reduces data accuracy).

Secure aggregation: When training models or computing results from many users, we use cryptographic protocols so the server only sees aggregated sums or averages, not individual contributions. For instance, federated learning with secure aggregation means the central server gets an update that is the sum of gradients from 1000 users, but it can’t tell any single user’s gradient. This ensures even if the server is compromised or curious, it can’t single out one person’s data easily.

Data trust and de-identification: A separate “key” service, perhaps run by a nonprofit or committee including participant representatives, holds the mapping from random user IDs to actual identities/email addresses. Researchers only deal with anonymized IDs. If data is to be shared with other scientists, it goes through that trust which removes any

remaining quasi-identifiers. This means no one on our analysis team even *could* look up who a particular person is, which prevents many abuses.

Audit logs and kill switch: The system keeps a transparent log of data accesses and deletions. If a user hits delete, it not only wipes their data (with cryptographic erasure if possible) but also logs that “User X invoked deletion on date Y” so an auditor can verify compliance. The kill switch covers backups and derived data too – we design so that e.g., if their data was used in an aggregated model, that model can be retrained or adjusted to forget their contribution if they wish (within reason; if they allowed use up to publication, we can’t pull back a published result, but we can for future analyses).

These technical measures are challenging to implement but are crucial for trust. We want our participants to know it’s not just promises – it’s coded into the system. The hope is that such rigor yields high participation and lower missingness (people don’t drop out just because they got spooked by something).

Ultimately, this protects dignity (no one feels like a “lab rat” under hidden cameras) and also improves data quality (no systematic biases from dropouts or withheld info due to privacy fears).

13.5.5 Give value back (so it’s not extractive)

To avoid the feeling that we’re just extracting data, we ensure the participants get *something* out of it – insight, utility, community.

Some features to give personal and collective value:

Personal ledger view: Each participant can see their own life’s data through a private dashboard. We design it carefully so it’s gentle and not anxiety-provoking. For example, it might show a long-term trend of their mood (smoothed, not day-to-day jaggles), their sleep pattern, and a “menu breadth” index over time. Importantly, we avoid any judgmental tone: it’s not like a fitness app saying “you did bad this week.” Instead, it’s observational: “In the past year, your lowest moods tended to occur in February, and your activity variety was narrower during those times.” This can help them reflect or discuss with a therapist if they choose. But it’s emphasized that these are *their* eyes only unless they share it.

Actionable nudges (opt-in): If a participant wants, they can get small suggestions based on their data – always *aimed at widening channels or comfort, not generic wellness spam*. For example, if the system detects their sleep has been irregular and that coincides with worse mood, it might gently suggest a proven sleep hygiene tip, or a prompt like “Consider a wind-down routine tonight? (We noticed your best weeks of

mood followed consistent sleep).” Or if someone’s pain flare tag correlates with not reaching out socially, a nudge could be: “It’s been a tough day – maybe call a friend or treat yourself kindly tonight.” These are never advertisements and never coercive. And they’re fully optional – user can turn off suggestions entirely. The idea is to use the data to actually help participants feel better or more in control.

Community science features: We might run periodic “open science weeks” where we share some aggregated findings with participants (like “this week we’re looking at how weekends vs. weekdays affect mood – see the anonymized result and compare yourself if you want”). Participants could even be invited to pose questions (“do people sleep less before they feel really bad?”) that we can test in the data and report back (if methodologically sound). We can also provide interactive tools with *simulated* data so they can learn how the models work. Basically, involve them as collaborators who are curious about the science. Perhaps even voting on features to add or surveys to run – a form of participatory research.

These feedback loops create a sense that contributing data yields personal insight and helps knowledge for all. It’s not a black hole where data goes in and nothing comes out.

13.5.6 A week in the life (what participation feels like)

To illustrate, here’s what an average week might look like for a Tier B participant:

Morning routine: They get a gentle ping after waking: “How do you feel?” with a valence slider, and “Energy level?” slider. Takes 10 seconds. Maybe once a week, an extra question: “How far ahead does life feel today?” (horizon slider) – but not daily to avoid annoyance.

Throughout the day: The app quietly counts steps and monitors phone usage for social metrics. It does *not* bug them during work or such. There’s no geolocation or audio recording, so nothing obvious happens. Essentially, they go about life normally.

Evening reflection: Around bedtime, a 60-second check-in: “Any big event today?” (yes/no), “Pain above normal?” (yes/no), “Caregiving duties today?” (yes/no). If they remember a dream from last night, a prompt: “Last night’s dream was: Pleasant / Neutral / Unpleasant / Don’t remember”, and maybe tag categories if they choose (“Stressful, included water, involved someone from childhood”, etc., from a list).

Sunday (or one chosen day): If they’re Tier B/C, once a week they play a short “horizon game.” For example, a simple game where they allocate effort between short-term and long-term rewards in a scenario (to gauge their horizon setting) – it’s designed like a little

challenge, not a dull survey. Or they fill a brief questionnaire about how they perceive future vs. present (time perspective inventory). We keep it around 3 minutes.

Passive data: The app and any wearables gather all needed signals passively. The participant doesn't have to do anything for those, just wear their device and carry their phone.

As-needed: If something like an fMRI is scheduled (Tier C), it's rare (annual) and coordinated with them with clear consent each time.

This schedule tries to embed into normal life with minimal disruption. Most days, it's two tiny interactions (morning and night) and otherwise just live your life. We want it to be as unintrusive as possible, so missing data is mostly because of life's ups and downs, not because the protocol is too irritating.

13.5.7 Scientific spine (so the data are actually useful)

To make sure this citizen science effort yields publishable, rigorous findings, we design the "spine" of the study with key endpoints and controls:

Pre-registered endpoints: We don't just let data swamp us; we decide on specific tests in advance even in the public setting. For example, we say: "Our primary endpoints are variance compression in end-of-life contexts (which mostly Tier C hospice participants can provide), repair-tilt after ledger shocks (Chapter 13.3's patterns), and horizon × menu interactions (like short horizon days showing lower entropy)". These are measured via clearly defined metrics and are the focus of analysis. The citizen science data is thus geared to answer those, not just an exploratory free-for-all.

Built-in negative controls: We incorporate things like randomized reminder timing (some users get check-in prompts at random times not linked to their state, to ensure prompt timing isn't influencing results) and "placebo event dates" (we might assign a fake 'event' date for each person far in the past and see if our pipeline ever picks up a pattern when nothing happened). These help confirm our analysis isn't finding false positives.

Replicable pipelines: We maintain a public repository from early on with analysis code, simulation testbeds, and synthetic datasets. Any derived metrics or processing steps are documented. This allows others (and participants, if they are savvy) to replicate our processing on fake data to see how it works. When we do have results, we can release aggregated data or code so that independent analysts could reproduce key figures on the anonymized dataset.

This spine ensures we *use* the citizen data effectively. Often, citizen science fails because lots of data comes in without a clear plan to analyze it. We avoid that by mapping data to specific falsifiable predictions of LoF.

13.5.8 Governance that deserves the name

Given the sensitivity and scope of this project, we need governance beyond just our research team:

Community IRB: We establish a standing Institutional Review Board or ethics panel that includes not just scientists, but participants (i.e., actual citizen scientists), clinicians (for the health aspects), ethicists, and statisticians. This board reviews the protocol initially and also any major changes (like if we wanted to add a new sensor or questionnaire later). They also watch out for mission creep or anything that might compromise ethics.

Red-team panel: We periodically invite independent “red teams” – basically skeptical experts – to stress-test our assumptions and analyses. For example, we might have a team of statisticians try to see if adaptation alone could explain our findings (the kind of rival theories we discuss in Part VII), or privacy experts to try to find any vulnerabilities in our data handling. They then publish their critiques (publicly or to the board). This keeps us honest and sharpens the science.

Public changelog: Every time we update the app, the model, or our analysis plan, we publish a versioned changelog that explains in plain language what changed and why. Participants can subscribe to these updates. For instance, “Version 2.1: Added a question about daily stress – after community vote – to better capture short-term shocks. Adjusted differential privacy noise on message counts to improve accuracy (ϵ increased from 2 to 3).” This way, there’s no sneaky changing of terms or methods behind the scenes. It’s all documented.

This governance approach aims to build credibility externally and trust internally. Participants see that it’s not arbitrary or exploitative; independent folks are watching the watchers.

13.5.9 Incentives that do not distort behavior

We want to encourage participation but not in a way that changes what we measure (the Hawthorne effect or perverse incentives could ruin data).

So we design incentives carefully:

Compensation proportional to time (for heavy tasks): For Tier C folks doing lab studies, we pay them like any research participant for their time and effort, at a fair hourly rate or

with meaningful thank-yous like gift cards or donations in their name. Importantly, we avoid any competitive or tiered payments that might encourage faking data. It's straightforward: do the scheduled task, get reward. We do not pay for "better" data or specific outcomes.

No lotteries or big prizes: Research shows lottery incentives can backfire and also might overly excite or disappoint some participants (hedonic artifacts). We avoid those. Any monetary incentive is modest and consistent – the aim is to honor effort, not to gamify life experiences.

Recognition in aggregate: We might acknowledge the community of participants in publications or on a website ("Thank you to the 5000 citizen scientists...") but we don't do individual leaderboards (like "Joe contributed the most data!") – that could pressure people or distort behavior (someone might start over-reporting to climb a board). The only exception might be a gentle acknowledgement for Tier C "methods helpers" if they consent, like listing names in an appendix, but not ranking or anything.

No built-in "achievements" for behavior: For instance, we won't have an achievement like "Logged mood 30 days in a row!" because that might encourage someone to log when they otherwise wouldn't (like dragging themselves while sick to just tap the phone to keep the streak). We explicitly say: it's okay to skip, no negative consequence. The app might even *randomly skip* some prompts (planned missingness) to remove any streak psychology.

Study ≠ self-improvement program: We refrain from typical wellness app features like setting step goals or giving "badges" for mood improvement. Those could change what we're measuring (if we encourage them to do 10k steps, we may artificially tighten variance for reasons unrelated to LoF). We keep the focus: we are observing, not coaching (aside from optional nudges which are subtle). If participants want to set personal goals, fine, but the app itself doesn't set them by default.

In sum, incentives are about appreciation and enabling participation (e.g., covering their costs), not about pushing people to behave in a certain way that would mess with the phenomena we want to observe.

13.5.10 Handling minors, crises, and clinical red lines

Our project deals with well-being which touches mental health, so we need strict policies for vulnerable cases:

Adults only (for now): The current design is for age 18+ (or whatever local age of consent). Minors would need a whole different design with parental consent, additional

safeguards, perhaps different measures. We may extend in the future carefully, but initially we exclude minors. That's partly ethical (minors can't fully consent, and their data is extra sensitive) and partly practical (developmental differences would complicate analysis).

Crisis protocol: If a participant's self-report indicates severe risk (like they answer a standard question that implies suicidal ideation above a threshold, or perhaps we include an optional PHQ-9 depression screening that flags danger), the app will *not* call 911 or anything without prior consent, but it will immediately show them local crisis resources (hotline, etc.). At onboarding, we might ask if they want a trusted contact notified in extreme cases; if they opt in, then in an emergency we could alert that person. But nothing behind their back. No police or involuntary interventions triggered automatically – that often does more harm. We essentially say: "If you ever feel in crisis, here's help" upfront, and if we detect high distress we remind them of those resources.

No secret clinician alerts unless consented: If someone is in a clinical trial or under medical care, they might choose to share data with their doctor. But otherwise, we don't send data to doctors or family automatically. It's under the participant's control if they want to export a summary to discuss with a therapist, for instance.

Clinical walls: We make it clear that this is a research study and personal tracking tool, *not* a medical service. It's not diagnosing or treating anything. If they have a medical condition, participation shouldn't replace seeking care. We set expectations that, say, if someone stops taking their antidepressant, our study's suggestions (or lack thereof) are not medical advice. We might incorporate this into the consent form: "This app is for research and self-reflection, not therapy or medical care."

Data use in clinical settings: If a participant does want to use their data for health reasons, we facilitate it safely (perhaps giving them a button to generate a PDF summary they can give their doctor). But we don't do it for them to avoid unwanted disclosure.

Essentially, we design as an observational study in the wild, not as a therapy or urgent care system. If someone shows extreme signs, we route them to appropriate help, but we don't take on the role of health providers.

13.5.11 Success metrics (for science and for trust)

How will we know if this citizen science approach is working? We set metrics both scientific and trust-related:

Scientific success: Ultimately, do we see the LoF signatures we hypothesized *replicate* across the broad cohorts and sites? For example, does variance compression near end-

of-life show up in Tier C hospice participants in multiple locations? Do those effect sizes hold up under the missingness sensitivity analyses of 13.4?. We aim for pre-registered effects to be confirmed with appropriate confidence. Also, an internal metric: have we avoided p-hacking? (Yes, if we stick to our plan.)

Missingness survival: One key indicator is that our missing data methods didn't break the results. If our results are robust after all that, that's a success. We will present that as: e.g. "The compression effect remained at ~0.20 (± 0.05) even under worst-case missingness assumptions". Surviving sensitivity checks is a success criterion.

Trust metrics: More qualitatively, *retention* is a measure of trust. High retention without coercion indicates people are okay with the study. We'll monitor dropout rates and reasons (maybe ask exit survey why leaving). If many cite privacy concerns despite our efforts, that's on us to improve.

We expect "frequent pauses without guilt" – i.e., participants use that pause button responsibly, and when they resume, they don't feel bad or get punished. We could measure how many pause and later come back (if it's a healthy fraction, it means people trust they can step away and return).

Participant ratings: We'll occasionally survey participants about their experience – clarity of consent, comfort level, whether they feel in control. Hopefully, we get positive responses that they understand what's happening and feel respected.

Open replication: A huge success would be if independent researchers (not involved in our study) take our public synthetic dataset or the eventual released real aggregate data and *replicate key findings*. Perhaps they use our pipeline on a different sample or run their own analysis and see the same patterns (e.g., horizon vs. variance negative correlation). If our materials are truly open and comprehensive, others can validate or even contest the findings. Successful replication or even extension by others would strengthen LoF's credibility greatly.

We will define some of these formally. For instance, "target retention > 70% at 6 months" as a trust metric; or "at least X out of Y primary outcomes significant after correction" as a science metric. The idea is not to post-hoc declare victory but to have initial benchmarks to strive for.

13.5.12 What would tell us we've failed

We also watch for signs of failure in the citizen science approach, so we can stop or fix it:

Mass drop-offs due to feeling "watched": If participation drops steeply after initial novelty, especially with feedback like "I got creeped out," that's a failure. Or if we see lots

of pauses specifically right after bad days (meaning participants might not want to report when unhappy), that indicates our methods might not have built enough trust or it's too burdensome at their worst moments.

Biased subset only: If horizon and repair-tilt patterns only appear in a small, highly engaged subgroup (like the uber tech-savvy folks) and vanish when we do leave-one-cluster-out (as in, a handful of super-users were carrying the effect), that's a fail. It would mean we didn't truly get a broad sample, and LoF might not hold generally.

Red-team replication of rivals: If our invited skeptics (red-team) demonstrate that simpler explanations (like hedonic adaptation, or just social support theory) can recreate our findings *without* needing any fairness constraint, and we cannot rebut that with unique evidence, then our interpretation fails. It means all this could be explained by known theories and LoF adds nothing predictive. In that case, even if data collection succeeded, the law itself would be in question.

If any of these persist despite adjustments or more data, we would have to either revise our approach or concede that the citizen science project did not confirm LoF. We would communicate this transparently (not spin it).

In practice, if we found, say, that many people quit because of privacy worries, we'd go back to Section 13.5.1–13.5.4 and enhance those measures or reduce data collected. If patterns only showed for a subgroup, we'd examine if that subgroup had some unique trait (maybe LoF is conditional on something we missed; or maybe our sampling method was flawed).

Takeaway: A citizen-science program for the Law of Fairness can indeed gather powerful evidence *without* being creepy or exploitative. By collecting only what's needed, fiercely protecting privacy, returning value to participants, and being brutally honest about missingness and biases, we create a virtuous cycle: more people join and stay, which gives better data to test the theory, which we then share and scrutinize openly. The result (if successful) is a truly democratic test of a deep scientific question. With that foundation, we can now look candidly for Section 13.6 Fail patterns: Expanding Variance—the kind of population-level outcomes that, if observed, would challenge the idea of any fairness constraint at all.

13.6 Fail Patterns: Expanding Variance

Up to now, we've focused on what *should* happen if a fairness constraint operates – things like variance contracting and trajectories converging near the end. But what if reality shows the opposite? If the dispersion in experienced well-being widens over time, especially when we'd expect compensation to rein it in, that's a serious problem for LoF. This section names specific failure signatures – patterns of data that would indicate LoF is false or incomplete – and outlines how to detect them and interpret each.

13.6.1 The core prediction about spread

First, recall the core LoF prediction about variance: Over a lifetime, individual streams should not permanently fan out to extreme happiness or misery – but instead there should be a transition back toward an overall neutrality felt state by the death of mind, meaning not just average reversion but a narrowing of the distribution of cumulative outcomes.

In population terms: If LoF holds, the distribution of people's cumulative affect (their life ledger residuals) should show reversion and variance contraction as horizons shorten (people get older or closer to death).

A Fail pattern would be persistent or worsening dispersion: for instance, if the spread (standard deviation or interquartile range) of lifetime well-being scores grows with age or illness rather than shrinks. Or if some lives just keep getting more and more outlier-level unhappy (or happy) without re-centering.

We quantify this by tracking variance of some summary like long-window average affect $\bar{x}_{w^{(i)}}$ for person i over a window w (like a year). Then see how $\text{Var}(\bar{x}_{w^{(i)}})$ changes across age or inverse horizon H^{-1} . We test the slope of variance vs. H^{-1} : LoF predicts it < 0 (variance down as horizons shrink). A failure is if we estimate $d(\text{Var})/d(H^{-1}) \geq 0$ or significantly > 0 . We will use mixed-effects models regressing variance on H^{-1} (with random intercepts for site/person) to get that slope and even Bayes factors to compare models with negative slope vs. no slope. If we find a zero or positive slope robustly, LoF's main quantitative claim is undermined.

13.6.2 Runaway tails (left or right)

Description: This Fail pattern is about the tails of the distribution. A heavy left tail means a subset of individuals end up in chronically and extremely negative states (unrepaired suffering), beyond what others experience. A heavy right tail means some lucky group rides on sustained bliss that never gets counterbalanced.

If LoF is true, such “runaway” cases shouldn’t persist generation after generation. Durable tails imply something is leaking – the system isn’t compensating those outliers.

Why it matters: LoF implies *compensability*, i.e. that extremely bad or good runs trigger corrections. If we see a stable subpopulation that remains in extreme misery, that suggests the system allowed a sustained imbalance (the left-tail leakage). Same for the right tail: if some consistently high-happiness people never come down to earth, maybe life isn’t forcing balance.

Tests: We’d examine the statistical tails of lifetime outcomes:

Use Hill’s estimator or other tail index measures on the distribution of net lifetime affect (or long-term residuals). If the tail index indicates a heavy tail (power-law-like rather than thin exponential), and especially if tail weight increases as horizon shortens, that’s bad.

Fit a Generalized Pareto Distribution (GPD) to top/bottom X% of the distribution and see if its shape parameter suggests heavy tails. Then compare these across age groups (are tails heavier among older folks or end-of-life groups? If yes, dispersion increased).

Specifically check if as people age, the worst-off 5% get relatively worse (further from the median) rather than closer.

Fail if: Tail heaviness increases with shrinking horizon or over time, given accessible channels. For example, if at younger ages the happiness distribution had mild tails, but by old age there’s a fatter left tail of very unhappy individuals that compensation didn’t rescue – LoF would be in trouble.

13.6.3 Horizon–variance decoupling

Description: This is when variance and horizon length simply have no relationship when they should. LoF expects variance to drop when horizons shrink (like nearing end of life). Horizon–variance decoupling means: take a group of individuals known to have objectively short horizons (say advanced cancer patients with 6-month prognoses) and another group with long horizons (healthy young adults), and after controlling for other factors, their affect variability is similar or even higher in the short-horizon group.

In other words, the supposed coupling (short horizon → less spread in mood/actions) isn’t observed.

What to check: We look for an interaction between horizon and channel access in predicting within-person HCl variability. Because one excuse for not seeing compression could be that channels were blocked (e.g., someone’s horizon is short but they’re also heavily medicated or isolated, so QS couldn’t do its thing). So we check: do those with

short horizons *and* intact channels show reduced variance? If not, that's a strong refutation.

Specifically:

Fit a model: $SD_HCl^{(i)} = \alpha + \beta_1 H_i^{-1} + \beta_2 ChannelAccess_i + \beta_3 (H_i^{-1} \times ChannelAccess_i) + \dots$
We preregister expecting β_1 negative, and especially β_3 negative (meaning the inverse-horizon effect is more pronounced when channels are open).

Fail if β_3 is near zero or positive and consistently so across cohorts. That would say even when channels are available, short horizon folks didn't show less variance than others.

Also more directly: take people who had clearly shrinking horizons (maybe their own report of "I feel time is short") and see if their variability in affect or behavior actually contracted. If we find no difference or even an increase, it's a decoupling.

This would mean LoF's horizon mechanism isn't functioning; maybe any observed balancing was just because of other reasons (like everyone just adapts regardless of horizon).

13.6.4 Bimodality without mixing

Description: Suppose we find the population separates into two clusters – some consistently high well-being, some consistently low – and individuals rarely move between those clusters over long periods. That's stable bimodality. If someone starts in the "unhappy" group, they stay there; if in "happy" group, they stay there, with minimal crossover or mixing.

Why it matters: LoF suggests there's a sort of "corridor" or attraction to neutrality. If instead, people settle into two (or multiple) distinct equilibria and don't converge, that contradicts a universal balancing force. It would imply maybe there are different "basins" of well-being that don't communicate (like chronic depression vs. not, and no compensation to pull you out if you're in the wrong basin).

Tests:

Compute Hartigan's dip test or Silverman's test on the distribution of lifetime outcomes or long-term average happiness to detect multimodality. If significant, distribution might be bimodal.

Use a Hidden Markov Model or similar on individual trajectories: do people flip states or mostly stay in one? If we find that once someone is in low well-being state for a while, they have a very low probability of transitioning out, that's problematic.

Look at mode residence times: how long individuals stay in a high or low affect regime. LoF would expect that even if they dip low, something eventually brings them out (residence time finite). If we see extremely long residence times that don't shorten when horizons shorten, trouble.

Fail criterion: If we confirm a persistent bimodal split that doesn't dissolve over time or with horizon changes. E.g., suppose older cohorts still show two separate happiness clusters with no narrowing compared to mid-life cohorts – that would suggest no converging force.

13.6.5 Widening cross-site dispersion

Description: If LoF is a fundamental law, we'd expect it to hold globally (unless blocked). If we find that in some cultures or sites people systematically end up happier and others systematically sadder *and that gap increases*, it could mean fairness isn't universal.

For example, imagine we track multiple countries: if over 50 years the variance between countries' average well-being increased (even after accounting for similar resources), it might suggest fairness isn't playing out globally, maybe due to sociocultural factors or it's not a true law.

Interpretation: LoF should be universal, not dependent on locale (assuming similar channel access). If cross-site differences grow uncontrolled, maybe fairness is not a law but context-dependent.

We check with hierarchical models:

Let each site/culture have a random mean and maybe random slope. Does the variance of site means increase over time? If yes, sites diverging.

Are some cultures always high well-being and continue to rise while others stagnate? That might mean some systems circumvent compensation (or perhaps suppress suffering differently, which would be a rival explanation like cultural baseline differences).

We have to control for measurement invariance across cultures. So we only consider this if we can put scores on same scale. Assuming that, we test:

If measurement invariance holds and we still see site-level SD of happiness growing as cohorts age, that's an issue.

Fail if: After controlling basics (GDP, etc. – basically giving LoF a fair shot by equalizing resources), the disparity in net outcomes between comparable sites increases with time

or across generations. That would imply fairness might be a local phenomenon or that unaccounted confounds (like politics) override any global constraint.

13.6.6 Intervention paradoxes

Description: We try a channel-widening intervention expecting it to help balance (like providing pain relief or social support to a group). If LoF is real, giving more channels should *reduce* variance and help net outcomes. A paradox failure is if such interventions do nothing or even make dispersion worse.

For example, a stepped-wedge trial where we give half a sample improved sleep tools and half later: if after the intervention, those with more open channels don't show any reduction in mood variance or left-tail misery compared to control, or weirdly they show more dispersion, that challenges the idea that channels facilitate compensation.

Design: Use stepped-wedge or randomized rollout of known relief measures (sleep therapy, access to counseling, etc.). Measure if variance or tail metrics change post-intervention relative to control.

LoF expects variance down or at least improved left-tail (less severe negatives).

Fail if: Across replications, we see that giving more relief opportunities *fails to change* dispersion or even correlates with *higher* dispersion. If even when we "widen channels" nothing happens to fairness metrics, maybe the whole mechanism is wrong or other factors dominate.

We also check consistency: if one study fails but others succeed, it might be context or power. But if repeatedly interventions that should relieve suffering don't bring distributions closer together, LoF's promise is questionable.

13.6.7 Ledger shock asymmetry that lingers

Description: We predicted asymmetry (13.3.8) as a sign of QS. A failure would be if negative shocks lead to long-term increased spread (people hit by bad events just stay worse off with no rebound) and positive shocks lead to no subsequent narrowing or counterbalancing.

In effect, after big hits, the ledger never recovers (persistent divergence), and after big gains, nothing eventually comes to pull them down to parity – they just keep that advantage or even widen it.

We test using event-study impulse responses (13.3.3 style):

If a large negative shock (like losing a spouse) yields a cumulative impact on ledger that stays significantly negative with no trend back up even after a long time (and confidence interval excludes zero compensation), that's a fail sign. Especially if we have channels open and still see no recovery.

If a large positive shock (lottery win) shows no eventual drag or increased generosity (no offset), just a permanently higher ledger, that also fails fairness.

So:

Compute cumulative impulse response of HCI or ledger after events. See if it ever approaches baseline.

Fail if: for negative events, even long-run IRF stays below zero (no rebound) and for positive events stays above zero (no correction), with statistically tight bands.

We'll look at confidence bands: if for negative shocks, the 95% band 2 years out is entirely below baseline (no recovery), that indicates a violation of presumed compensation.

13.6.8 Social network amplification

Description: Perhaps networks cause a Matthew effect (the rich get richer, poor get poorer in affect). Maybe people who are well-connected accumulate more relief (friends help them more), while isolated individuals accumulate more strain. If this effect grows, networks could amplify inequality of well-being.

Why it matters: LoF would ideally operate *within* any network structure – it shouldn't allow persistent unfairness just because of social topology. If we find that hub individuals (many friends) systematically end up way better off and peripheral individuals worse off, and that gap increases, then LoF might not be an overarching constraint, or only works if networks are egalitarian.

We analyze:

Use graph-aware mixed models: include measures like degree (how connected someone is) and see if the Gini coefficient of well-being across degree groups increases.

Compute Lorenz curves of affect by centrality quantiles. If the Lorenz curve bends more over time (meaning higher inequality associated with network position), that's not what LoF would predict.

See if high-degree nodes accumulate disproportionately positive ledger changes, while low-degree accumulate negative, beyond what resource differences explain.

Fail if: After controlling resources, we still see increasing Gini of affect by network degree over time. In plainer terms, if being popular leads to feeling better and better relative to loners and LoF doesn't level that out at end-of-life or via internal mechanisms, then fairness might be more a social artifact than law.

This could indicate that perhaps LoF only works if network effects are accounted for, or that networks themselves are part of the mechanism in a way we didn't think (this might tie into social baseline theory or such, which would be a rival explanation overshadowing LoF).

13.6.9 Dream counterweight failure

Description: Chapter 10 posits that dreams often invert the emotional tone of the prior day as a low-cost counterweight. A Fail pattern would be if this doesn't hold – after tough days, dreams are *not* lighter or more positive than day (no counterweight), and maybe even reinforce negativity (nightmares adding insult to injury). If variance *expands* across day and night (bad day followed by bad dream, etc.), that's against the idea of nightly relief.

Test:

In participants who report dream affect and daytime affect, check correlation: is it negative (as predicted) or zero/positive?

We might do controlled lab studies (like Chapter 10 tasks) with awakenings after stress days. If those don't show compensatory dream effects across many people/labs, that's a fail.

If REM-sampled dreams on difficult days don't show the predicted positive shift (or threat simulation that helps adjust), then one channel of QS (dreams) might not be working.

Fail if: In within-person analyses, *no* inversion of dream vs. prior day affect is found, even under conditions we expected it. And if we replicate that across labs: e.g., after trauma, dreams are just as negative as day or even worse, contrary to expectation.

This would remove a hypothesized major “no-cost” compensation channel, thereby weakening LoF because then real-world compensation would have to do all the work.

13.6.10 End-of-life dispersion persists

Description: The ultimate test: near the very end, do people's experiences converge or not? Fail pattern here is if in hospice or last weeks (with mind intact and channels open), we see *no* variance compression and *no* tilt towards repair; in fact, maybe some people

get even more divergent (some suffer terribly, others not, and no balancing for those who suffer).

If with all our ideal conditions (pain managed, support given) the end-of-life data still shows wide outcome variability and no trend to neutral, LoF pretty much fails at its sharpest test.

We plan:

Analyze hospice cohorts: measure SD and left-tail of affect in, say, final month vs. 6 months prior. If not lower, that's a fail.

Check composition: if some patients continue adversarial or unresolved behaviors to the end, and across many cases not balanced by others, that's again not compressing.

Specifically, we'll look for:

Variation (SD) of daily HCl in the last week of life compared to a month earlier. LoF expects down; fail if not.

Distribution of final ledger residuals among those who died with open channels: do many end significantly negative or positive beyond equivalence margins (Ch. 11 suggests \pm some HCU margin)? If yes in a systematic way, fail.

Fail if: End-of-life windows show no drop in variance or tilt relative to earlier periods, after controlling sedation etc. Also if final ledgers often fall outside our predefined “neutral range” and not just by tiny amounts, even when conditions for fairness were seemingly met, that basically falsifies the Law.

If we find, say, 30% of hospice patients ended with very negative ledgers despite all care (and more than would be expected by chance), then LoF is likely false or incomplete.

13.6.11 Statistical dashboard (what to plot)

We have a slew of tests, and to monitor them, we'll maintain a dashboard of plots for visual checks:

Variance vs. horizon curves: For each cohort or site, plot mean variance of affect against horizon (age or time to death) with credible bands. We expect downward trend. If we see flat or upward in any, highlight it.

Tail mass metrics over time: Plot proportion of individuals with extremely negative outcomes (<10th percentile) and extremely positive (>90th) across age.

Bimodality diagnostics: perhaps density plots by age group to see if distribution becomes bimodal or not mixing.

Gini/Atkinson index over time or by network degree quantile: if rising, visible.

Impulse response functions (IRFs) for events: overlay for different events to see if they revert to zero or not.

Intervention step plots: Before vs. after dispersion for those got channel widening vs. not.

All with negative controls (like horizon permuted or random grouping) to ensure we're not seeing spurious patterns.

This dashboard will be regularly updated as data comes in, so we can spot any emerging Fail patterns early and adjust analysis or at least know where LoF is weakest.

13.6.12 Interpretation chart (quick read)

To make it easier for readers, we create an “interpretation cheat sheet” for these patterns:

If we see single-site variance up but cross-site stable and horizon-linked, likely a measurement or channel-blockage issue -> fix design (maybe that site had poor data or lacked an intervention).

If cross-site variance up and horizon-unlinked, that hints the law might not be truly universal (model failure or LoF not global).

If tails thicken and bimodality persists everywhere, that's strong disconfirmation of LoF (the worst-case).

If variance contracts with horizon, tails thin, interventions help across contexts, that supports LoF is active.

This chart basically guides how we'd react:

Some Fail patterns (like small deviance in one site) might mean just refine methodology.

Others (like persistent heavy tails, or no end-of-life compression anywhere) mean go back to drawing board theory-wise.

13.6.13 Minimum bar for claiming success

We want to be strict: we won't declare LoF “validated” unless it clears a high bar. At minimum:

We see a negative horizon-variance slope replicated in independent groups/labs (not just one p-value in one dataset).

We see tail thinning and state mixing near short horizons (i.e., evidence that extreme states are moderated as the end nears, and people don't get stuck permanently in one extreme).

Interventions that widen channels reduce dispersion in a measurable way (e.g., those given better pain relief show less variance or quicker rebounds than those without).

And crucially, these findings persist after all our MNAR and sensitivity analyses, leave-one-out checks, etc. (no fragile, one-analyst special effects).

If rivals can explain the findings, we haven't reached the bar.

If those are all met, then we feel comfortable saying the evidence supports LoF as a law-like constraint. Anything less, and we couch it as at best a tendency.

Takeaway: Fairness, if real, should reveal itself not just in average trends but in the very shape of lived experience. When time runs short and chances to balance are present, we should see lives not flying apart in chaos but narrowing in spread, thinning out the extremes, and coming back toward the middle. If instead we observe the opposite – widening gaps, entrenched misery vs. joy camps, no effect of giving people help – then either our measurement is flawed or the Law of Fairness is not governing our reality. This final check is decisive: either the patterns of dispersion support a compensatory law or they tell us that what we've been seeing was coincidental.

13.6.14 Where we go next:

Next, in Part VII – Rival Explanations, Fairly Presented, we subject the Law of Fairness to its greatest tests: comparing it head-to-head with alternative theories (hedonic adaptation, predictive coding, etc.) to see if those frameworks can also (or better) explain the patterns we've found. If a simpler theory accounts for everything, LoF doesn't earn its keep; if not, LoF gains credibility as something genuinely new.

Part VII — Rival Explanations, Fairly Presented

After laying out the Law of Fairness in preceding chapters, we now turn a critical eye toward alternative explanations. Part VII serves as a rigorous stress test for our theory, adopting a scientific stance that demands every claim be challenged by plausible rivals. A core principle of science is that any hypothesis – no matter how compelling – must compete against rival explanations under fair conditions. In this spirit, we will interrogate other theories that could explain the same fairness phenomena our Law of Fairness purports to cover. By doing so, we ensure that our conclusions are not born of tunnel vision or confirmation bias, but stand strong against informed critique.

Our approach in this part is to present each alternative account on its own terms, evaluating it with precision and impartiality. We refrain from straw-man arguments or cursory dismissals; instead, each rival hypothesis is given a fair hearing with attention to both supporting evidence and counter-evidence. The goal is to identify conditions under which a given rival might outperform the Law of Fairness in explaining human behavior, as well as contexts where it fails to account for critical observations. This balanced analysis embodies a human-first perspective: we care about which theory best explains real human decision-making, not just which is philosophically pleasing or convenient. Throughout, testable predictions remain front and center — each explanation is probed for how well its predictions match empirical reality.

In scope, Part VII canvasses alternatives spanning from evolutionary biology to economics and cultural anthropology. We will examine hypotheses that attribute fairness to genetic self-interest, strategic reciprocity, social learning, or even statistical flukes, among others. Each chapter in this part shines light on one or more of these competing ideas. Importantly, by confronting these alternatives head-on, we do more than critique them – we also clarify the boundaries of validity for our own Law of Fairness. If a rival theory better explains certain fairness behaviors, that reveals where our law must be refined; if it falters where our law succeeds, that reinforces the unique explanatory power of the Law of Fairness. In sum, Part VII is about putting our theory to the test and ensuring that, in the landscape of possible explanations, the Law of Fairness stands not by default, but because it has proven to be the most robust and comprehensive account of the evidence.

What you'll get from this Part:

- Rival theories tested head-to-head: See how each alternative explanation for fairness (from hedonic adaptation to social learning) is presented at its strongest and evaluated on equal footing with the Law of Fairness.

- Clear criteria for success: Learn about five key empirical signatures of fairness (horizon effects, end-of-life neutrality, dream counterweights, Queue System residuals, and population-level balance) that we use as yardsticks to judge whether a rival theory can match LoF.
- Strengths and limits of each explanation: Discover where each competing theory shines (e.g. kin selection for family altruism, reciprocity for long-term cooperation) and where it falls short (e.g. failing to explain one-off kindness to strangers or terminal emotional patterns).
- A sharper or revised LoF: Find out whether confronting these alternatives ultimately bolsters the Law of Fairness as uniquely comprehensive or reveals specific gaps where our theory must be improved. Either way, the exercise ensures no claim survives without scrutiny.

What counts as a serious rival? We treat as serious only those frameworks that provide:

- Mechanism – story linking brains, bodies, and behavior, not just a vague principle.
- Scope – coverage across life, dreams, environments, and end-of-life phenomena.
- Specificity – quantitative predictions (not just post-hoc narratives).
- Falsifiability – clear signatures that would count against the theory.
- Comparative fit – ability to match or beat LoF on preregistered tests.

The short list we will examine:

- Hedonic Adaptation / Set-Point Theories. People drift back toward a baseline level of happiness after gains or losses.
- Opponent-Process Theory. Every intense affective state triggers a compensatory opposite state over time.
- Predictive Coding / Free-Energy Principle (FEP). Organisms minimize variational free energy (a bound on surprise); affect reflects precision-weighted prediction errors.
- Reinforcement Learning with Homeostatic Control. Agents maximize rewards over time subject to internal set-points and costs of deviation.
- Social Baseline and Attachment Accounts. Social proximity reduces metabolic and emotional costs; isolation increases them, shaping affect within lifespan.

- Stochastic “Good/Bad Luck” Mixtures. Affective life is mostly noise plus regression to the mean, with no deeper constraint.
- Composite Hybrids. Blends of adaptation, predictive coding, homeostatic, and social mechanisms that claim to recover all observable patterns.

Each of these has explanatory power. Our task is to ask whether they can also account for the LoF signatures without the central commitment that cumulative felt experience is constrained to neutral at the death of mind. Throughout the next chapters we will return to five key discriminators. A rival theory earns points if it predicts these patterns *a priori* and loses points if it must add post-hoc patches to explain them.

What this Part will do for you:

- Explain Horizon interaction. As subjective time runs short, choice menus narrow and tilt toward reparative or relief-focused actions, and variance in affective outcomes contracts. LoF expects a strong effect of the remaining horizon on the “admissible set” of actions. Rivals must show the same horizon-specific narrowing arises from their own principles, without invoking any external fairness mandate.
- End-of-life neutrality. When cognition is intact and social/psychological channels are open, affective dispersion markedly narrows near death; reconciliations and relief acts cluster and the net “ledger” moves toward neutral. LoF predicts measurable variance compression and near-neutral cumulative affect at life’s end (illustrative preregistered targets: mean within ± 0.15 z, slope within ± 0.05 z/day; variance ratio ≤ 0.80 , where z denotes standardized units on the chosen affect scale). Rivals must show equal or better fit using only adaptation, free-energy, or social mechanisms, with no built-in push toward a zero-sum ledger.
- Dream counterweights. After emotionally difficult days, dream affect tends to invert or lighten; over a week, sleep appears to contribute to rebalancing the ledger (e.g. an intense negative day is followed by calmer or positive dream content). LoF expects systematic counterweight dynamics in dreams. Rivals must show the same signature can emerge from standard memory consolidation or prediction-error “clean-up” alone (e.g. generic rebound or stress processing), without any special balancing drive.
- QS-like residuals in control hubs. After accounting for utility, conflict, arousal, and risk, there remains a systematic “menu-shaping” signal in brain regions like rIFG, ACC, vmPFC that tracks compensatory potential (as posited by the Queue System in Part III). LoF predicts a residual neural pattern tied to horizon and repair-tilt (Φ).

Rivals must show that control or precision-weighting processes would produce these residual patterns on their own, rather than requiring a new term.

- Population-level dispersion bounds. Over years and decades, when channels for adjustment are intact, the spread of cumulative affect across individuals is bounded; the most extreme life outcomes (“tails” of total happiness or suffering) are systematically thinned as horizons shrink. LoF expects bounded dispersion and tail-thinning in lifetime well-being distributions. Rivals must demonstrate that the same population bounds follow from known adaptive, network, or homeostatic forces, rather than an overarching fairness constraint.

A fair playing field: To avoid moving goalposts, we will grant each rival its strongest plausible form. That means, for example, allowing adaptation models to have multiple time constants and context-sensitive gating; giving opponent-process models asymmetric on/off dynamics; letting predictive coding (FEP) models include precision control, flexible policy selection, and interoceptive predictions; augmenting RL models with homeostatic set-points, risk sensitivity, and social rewards; and extending social-baseline models with cultural and lifespan factors. Where necessary, we will even suggest the cleanest experimental tests that would allow a rival to shine. If a rival still cannot produce horizon-dependent narrowing, terminal variance compression, or QS-like residuals without extra assumptions, that weighs in LoF’s favor.

Chapters in this Part:

- **Chapter 14 — The Hedonic Treadmill and Opponent Processes** - revisits hedonic adaptation and opponent-process theory. We will see where they succeed – short- and medium-term emotional rebounds – and where they struggle, such as explaining horizon-sensitive choices and strict end-of-life neutrality.
- **Chapter 15 — Predictive Coding and Free-Energy** - asks whether minimizing surprise and prediction error alone could impose fairness-like outcomes (or if it essentially becomes LoF once tuned to do so).
- **Chapter 16 — Reinforcement Learning with Homeostatic Regulation** - explores whether an agent with set-points, reward maximization, and physiological limits can recreate LoF’s signatures without baking in a fairness constraint. Each chapter presents the rival view as fairly as possible, then highlights specific tests that could distinguish it from LoF.

Where we go next:

Chapter 14 begins our head-to-head tests. We start with the strongest familiar accounts—hedonic adaptation and opponent processes—and ask whether their error-correction dynamics, on their best day, can produce the horizon-sensitive narrowing, terminal variance compression, and ledger-like counterweights that the Law of Fairness predicts without adding new assumptions.

Chapter 14 — The Hedonic Treadmill and Opponent Processes

We initiate our examination of competing theories by laying out the primary alternatives to the Law of Fairness and setting the stage for their evaluation. The specific purpose of this chapter is to frame how seemingly “fair” behaviors might be explained without invoking any special fairness principle – in other words, to ask whether our observations could be accounted for by more conventional or skeptical models. This inquiry is crucial: if common explanations suffice, then a dedicated Law of Fairness would be unnecessary. Thus, the chapter serves as a foundation for stress-testing our theory, articulating what each rival hypothesis claims and what it would mean if that hypothesis were correct.

The reader will see clearly what is at stake: each alternative, if valid, would fundamentally challenge the need for a Law of Fairness, so we must understand them in order to properly assess our theory’s strength.

What you’ll get from this Chapter:

- Major affect-regulation explanations, side by side: A nuanced tour of hedonic adaptation and opponent-process theory, showing how each attempts to account for apparent balancing without invoking a special law.
- Strengths and shortcomings highlighted: Clear insight into what each rival theory explains well (e.g. rebounds after gains/losses; after-reactions to intense states) and where each falls short (e.g. guaranteeing horizon-dependent menu shaping or terminal variance compression).
- Why (and where) LoF stands apart: An appreciation of why the Law of Fairness still emerges as necessary after examining all contenders – or precisely which gaps in our theory the rivals reveal, ensuring no aspect of fairness is taken for granted.

Subsections in this Chapter

- **14.1 Best Evidence For** - Presents the strongest evidence for hedonic adaptation and opponent-process theory: classic rebounds after gains and losses, short- and mid-term moderation toward baseline, and well-described biological mechanisms. Establishes the best case these rivals can make without straw men.
- **14.2 Where They Shine** - Identifies contexts and timescales where these accounts fit the data well (e.g., post-event mood normalization, pharmacologic/opponent after-effects, everyday return to set-points) and clarifies the boundaries of those successes.

- **14.3 What They Cannot Guarantee** - Shows what remains unproven under “tendency” models: no per-life guarantee of a neutral ledger, no required horizon-dependent intensification near the end, and no prediction of terminal variance compression. Notes that extreme or asymmetric environments can leave lasting imbalance under these rivals.
- **14.4 Research Notes: Tendency vs. Law** - Formalizes the distinction between statistical tendency and a law-like constraint. Lays out preregistered tests (e.g., horizon effects, end-of-life variance compression) and an out-of-sample metric to adjudicate models, guarding against regression-to-the-mean and optional-stopping artifacts.

Where we go next:

We now turn to Subsection 14.1, where we present the best evidence for hedonic adaptation and opponent-process theory in their strongest form, setting the baseline these rivals must meet before we test where they fail to capture LoF’s signatures.

14.1 Best Evidence For

If you've ever felt a promotion's thrill fade into routine, or watched heartbreak's agony mellow into a dull ache, you already grasp the intuition behind the hedonic treadmill. It is the idea that our emotional state tends to return to baseline after both triumphs and tragedies. Decades of research back this up. In a classic study, *Brickman (1978) compared lottery jackpot winners with people who had been paralyzed in catastrophic accidents*. Remarkably, after some time had passed, neither group was as far from "normal" happiness as one might expect: the lottery winners were not ecstatically happy for life, and the paraplegic individuals were not perpetually despondent. In fact, in that sample both groups' happiness levels drifted back toward where they had started. This finding, though initially surprising, has been replicated and nuanced by large longitudinal studies. Major life events (marriage, divorce, job loss, widowhood, etc.) do shift happiness in the short run, but *in general* people tend to regress toward their own mean happiness level over time. The adaptation doesn't happen overnight, and it isn't always 100% complete, but the tendency to rebound is one of the most robust observations in well-being research. There is even evidence that individuals have a *set-point range* for happiness—partly heritable and linked to personality—that acts like a thermostat, pulling mood back to a characteristic range after disturbances. All of this is strong evidence *for* the view that some self-regulating mechanism (psychological or physiological) works to stabilize our affective life.

Another line of evidence comes from opponent-process theory, originally proposed by Solomon and Corbit in the 1970s to describe the dynamics of emotional reactions. The core idea is that every primary emotional state is automatically followed by a contrasting "opponent" state that moderates it. The classic example is *fear and relief*. When a novice skydiver jumps out of a plane, they feel an intense surge of fear on the way down and only mild relief upon landing. But as they gain experience with repeated jumps, the primary fear response weakens, and the after-response of euphoric relief grows much stronger. In one study, first-time skydivers reported extreme terror and minimal pleasure after landing, whereas experienced skydivers reported little fear during the fall and a *pronounced* wave of pleasure upon safe landing. Similarly, opponent-process dynamics are evident in drug use and other stimuli: an initial positive high is followed by an opposing crash or withdrawal, and over time the positive effects diminish while the negative after-effects strengthen. Even everyday experiences show this pattern — for instance, the sweet relief after pain: immersing your hand in ice water causes pain, but shortly after you remove it, a pleasant soothing sensation arises as the opponent process. Notably, the nervous system itself enforces such rebounds. Sustained pain, for example, can recruit descending analgesic pathways (including endogenous opioid

mechanisms) that diminish pain perception (Fields, 2004), illustrating an innate neural circuit for restoring equilibrium. These examples underscore that the body and brain seem wired to push back against prolonged extremes. Intense activation of one emotional polarity (pleasure or displeasure) triggers processes that eventually produce the opposite sensation, bringing you back toward an emotional middle ground.

In summary, the best evidence supporting these rival theories comes from a wide range of human (and animal) studies demonstrating *reversion toward the mean* and *after-effects*. People adapt to both positive and negative life changes to a significant degree, and strong emotional reactions are countered by opposite reactions that help restore equilibrium. These findings form a formidable foundation: any theory of life's emotional trajectory must, at the very least, be consistent with the fact that *most* ups and downs are temporary. LoF does not dispute any of this evidence – on the contrary, it inherits these well-established phenomena as part of what must be explained. The question is whether these processes, by themselves, are enough to account for *all* the patterns we observe (especially the more subtle, long-horizon ones), or whether an additional law-like constraint is needed.

14.1.1 Longitudinal rebounds after major life events

- Life shocks normalize. Decades of panel data and cohort studies show that emotional impact from marriage, divorce, childbirth, unemployment, disability, and widowhood tends to attenuate over months to a few years. Trajectories typically display a sharp initial displacement followed by partial or substantial return toward the person's pre-event level.
- Strength of evidence. Large N, repeated measures, within-person designs, and fixed-effects models reduce confounds (stable traits, cohort differences). Many studies also document anticipation effects (drift before the event) and asymmetries (losses adapt more slowly than gains), suggesting genuine dynamics rather than mere reporting bias.
- Boundary conditions. Adaptation is incomplete for some events (e.g., severe disability, chronic pain) and can be slower in contexts with sustained stressors or minimal social support. These are not counter-examples to adaptation; they refine where and how fast it operates.

14.1.2 Experience sampling and the “tether” to baseline

- Daily return. Smartphone and diary studies show that mood perturbed by hassles or uplifts tends to mean-revert within hours to days. Autoregressive models

reliably estimate negative feedback (yesterday's deviation predicts today's drift back).

- Robustness. Findings replicate across ages, cultures, and measurement instruments (Likert affect scales, linguistic sentiment, brief physiological markers), supporting a trait-like set-point plus state-dependent rubber band.

14.1.3 Psychophysiological after-effects (opponent dynamics)

- Classic pattern. Intense fear is followed by relief; acute pain by parasympathetic rebound; thrill by a subdued, sometimes “blue,” after-state. Startle, threat-imminence, and pain paradigms demonstrate overshoot and counter-swing in heart rate variability, skin conductance, and pupil response after the primary driver ends.
- Temporal specificity. The time course of the “b-process” (the opponent) grows with repeated exposure and outlasts the “a-process,” matching the original formalization of opponent-process theory.
- Behavioral signatures. The counter-swing influences choice: after relief, risk tolerance briefly rises; after strong positive arousal, people may show transient irritability or fatigue.

14.1.4 Neural recovery and baseline stabilization

- Return-to-set signatures. fMRI, intracranial recordings, and EEG show that limbic responses (e.g., amygdala) to affective stimuli dampen with repetition or time, while control/valuation regions (vmPFC, striatum) display adaptive recalibration—a neural correlate of “getting used to it.”
- Circuit motifs. Homeostatic inhibitory motifs (prefrontal–amygdala, anterior cingulate–periaqueductal gray) and neuromodulatory gain control (serotonergic and noradrenergic tone) implement negative feedback that curbs persistent extremes, consistent with both treadmill and opponent-process views.

14.1.5 Pharmacology and addiction: the hard test of opponents

- Tolerance and withdrawal. With repeated drug exposure, the initial positive effect shrinks (tolerance) and an opponent negative state (dysphoria, craving) strengthens and outlasts the drug—textbook opponent-process. Indeed, addiction research confirms this pattern: chronic drug use recruits an “anti-reward” stress system that induces persistent dysphoria (Koob & Le Moal, 2008). In other words, excessive artificial highs provoke the brain to swing the mood

pendulum back toward pain, a sobering demonstration of nature's compulsion toward affective balance.

- Allostasis. Chronic use appears to reset hedonic set-points downward, yet the acute counter-swing remains visible after each dose. This pattern is difficult to explain with simple regression-to-the-mean and strongly supports process-level opposition.

14.1.6 Sensory and motor analogies (opponent coding is ubiquitous)

- Vision and color after-images, balance (vestibular push-pull), and motor control all rely on opponent channels with well-characterized dynamics. While not affect per se, these systems demonstrate that biological control commonly uses paired opposites for stability—supporting plausibility for affective opponents.

14.1.7 Sleep and dreams as normalizers

- Overnight repair. Poor days can be followed by changes in REM sleep and emotion-laden dreams that blunt next-day reactivity; conversely, REM deprivation can exaggerate amygdala responses.
- Memory reconsolidation with affect damping. Sleep supports re-tagging emotional memories with less autonomic punch. Whether we label this “opponent” or “homeostatic normalization,” it is consistent with rebound frameworks.

14.1.8 Social baselines and buffering

- Attachment dampers. Social proximity and supportive contact reduce physiological load and speed affective recovery, producing faster returns to baseline after stressors.
- Network redundancy. People with richer social scaffolds show quicker mood normalization after shocks—compatible with treadmill dynamics mediated by social regulation.

14.1.9 Developmental and lifespan trends

- Age-related smoothing. Older adults often report lower affective volatility and quicker recovery from daily hassles, paralleling enhanced opponent-like control and strategic attention (positivity effect).

- Learning effects. Repeated exposure to classes of events (e.g., public speaking) yields smaller perturbations and faster rebounds—an experiential refinement of set-point and opponent mechanisms.

14.1.10 Quantitative models that fit the data

- Multi-timescale adaptation. Models with a fast and slow feedback term capture both quick dampening and gradual trait-level drift.
- Asymmetric opponents. Allowing the opponent process to grow with repetition and to decay more slowly than the primary process reproduces after-sensations and rebound aversion/relief.
- Context gating. Incorporating resource and social context improves fits to panel data without invoking new metaphysics: rebounds are conditional on ambient support and load.

14.1.11 Where this evidence is strongest methodologically

- Within-person designs with long follow-up minimize between-person confounds.
- Pre-registered ESM (experience sampling) with high compliance reveals short-horizon mean reversion.
- Physiology + behavior in the same sessions connects subjective rebound to objective counter-swing.
- Manipulations (sleep restriction/extension, acute stress, pharmacology) produce predictable changes in the size and speed of rebounds.

14.1.12 What these frameworks already explain—without extra assumptions

- Why “good” and “bad” do not last at full intensity.
- Why repeated exposure softens impact but can create after-effects that feel opposite.
- Why support and rest matter for quicker normalization.
- Why dreams often feel emotionally “busy” after intense days and calmer after ordinary ones.
- Why age and practice are associated with steadier affect.

14.1.13 Interim conclusion (in favor of the rivals)

If we stop here, the hedonic treadmill and opponent-process theory already offer a coherent, mechanistic, and well-evidenced account of much of emotional life. They require no global constraint and still explain everyday rebounds, physiological counter-swing, neural dampening, sleep-related normalization, and the role of social context. They also make testable, quantitative predictions that are borne out across multiple paradigms.

14.1.14 Where we go next:

Next we examine where these mechanisms shine across contexts and timescales. Section 14.2 surveys tasks and life stages where adaptation and opponent processes match the observed trajectories without extra machinery.

14.2 Where They Shine

Hedonic adaptation and opponent-process theory have risen to prominence for good reason: they elegantly explain *why* our emotional life isn't all runaway highs and spiraling lows. Let's consider their strengths in turn.

Adaptation (the Hedonic Treadmill) excels at explaining *medium-term stability*. We all know the story: get a new raise, a bigger house, or a shiny gadget, and you're happier for a while – but soon enough, the excitement fades and your mood returns to roughly where it was. Conversely, even after devastating losses or injuries, many people eventually find their footing again and no longer feel the intense grief or shock they initially did. Adaptation theory accounts for this by positing psychological mechanisms that *dampen the long-term impact of change*. These include processes like habituation (we simply get used to new circumstances) and cognitive shifts (we adjust our expectations and values). Empirically, there is strong evidence for such processes: people's aspirations rise with their incomes, so the happiness boost of more money is short-lived; attention shifts away from changes as they become the new normal; and memory biases cause past extreme events to lose their emotional sting over time. Biologically, adaptation is supported by phenomena such as neural desensitization – for instance, rewarding stimuli provoke less dopamine release after repeated exposure, preventing a permanent high. In short, adaptation is a built-in emotional *homeostat*: it explains how we maintain consistency and why emotional deviations are usually corrected over time.

Opponent-process theory shines in explaining *short-term dynamics and the pendulum-like nature of affect*. It provides a framework for why *every peak comes with an after-peak*. The fact that joy can turn into melancholy once the excitement passes, or that pain can give way to calm, is predicted by this theory's central premise: the brain has paired processes for emotion – one that swings one way, and another that swings back. This accounts for the *transient overshoot* phenomena: e.g., the post-exam crash after weeks of stress, or the post-performance blues that artists report after a big show. On the flip side, it accounts for the solace after a crisis: people often describe a strange serenity or clarity after a period of intense fear or grief, as if an emotional counterweight kicked in. Opponent-process mechanisms have clear physiological analogues. For example, the sympathetic nervous system triggers fight-or-flight responses (increasing heart rate, alertness, etc.), but once the threat is gone, the parasympathetic system pours on the brakes, leading to relief, fatigue, or contentment. In addiction research, opponent-process theory provides a compelling explanation for drug tolerance and withdrawal: the body fights the drug's effects with opposing reactions, so more drug is needed to achieve the same high, and when the drug is absent the opponent process (now unopposed)

creates distress. These examples show the breadth of behaviors the theory can cover – from thrill-seeking and exercise “highs” to the physiological basis of addiction and emotional regulation.

Together, adaptation and opponent processes cover a lot of ground. They explain why most people, most of the time, hover around moderate mood states (extremes are self-correcting), and why strong perturbations trigger corrective responses (every emotional action has a reaction). They have been validated in laboratory experiments, longitudinal studies, and everyday observations, and they align with known biological feedback systems. In terms of scientific virtues, these theories are parsimonious (no need to invoke mysterious forces – simple feedback loops suffice) and broadly applicable. They don’t depend on one’s philosophy or moral outlook; they emerge from basic properties of neural and psychological systems. As rivals to LoF, they are formidable because if life’s apparent emotional balancing can be fully explained as the aggregate result of these well-understood mechanisms, then there is no need to postulate a new “law.” In essence, hedonic adaptation and opponent processes *already* embody nature’s fairness to a large extent: they ensure that no pleasure or pain lasts indefinitely and that our feelings gravitate toward equilibrium. This is their zone of triumph, and any proponent of LoF must concede this: the rivals get a lot right.

To illustrate their explanatory power with a concrete case: imagine a person who goes through a breakup and is heartbroken, then one year later feels okay and even optimistic about life again. Adaptation theory straightforwardly says: time heals – through a combination of adjusting to daily life without the partner, cognitive reframing (“maybe it wasn’t meant to be”), and the natural waning of intense emotions, the person’s happiness returns to baseline. Opponent-process theory adds that during the initial aftermath, when grief was at its peak, the seeds of its opposite (a kind of relief or independence) were already being sown and grew over time, helping the person bounce back. We don’t need any cosmic ledger to understand this rebound; it emerges from normal human resilience and regulatory biology.

In summary, these rival explanations *shine* in accounting for short- to mid-term affective regulation. They set the baseline expectation that any new theory must at least match: life has a self-correcting emotional economy, driven by adaptation to change and opponent responses to extremes. LoF, if valid, would need to incorporate these dynamics, not contradict them. Indeed, one might say LoF takes these everyday balancing acts and asks, “Do they add up to something even more regular when we consider an entire life or system?” But before considering what LoF adds, we must fully

appreciate that the hedonic treadmill and opponent processes already explain a great deal about why emotional life stays within bounds.

14.2.1 Short-to-mid horizon regulation

- Fast rebounds after everyday perturbations. Hassles, uplifts, wins, and snags show mean reversion over hours to weeks. Autoregressive fits are stable; volatility shrinks as perturbations age.
- Context-contingent set-points. Baselines shift with sleep, illness, social support, and workload, then re-stabilize, matching multi-timescale homeostasis without invoking any ledger-like closure.
- Practice effects. Repeated exposure to public speaking, pain, or evaluative threat yields smaller perturbations and quicker recovery—classic adaptation.

14.2.2 Opponent after-reactions and allostasis

- Physiological counter-swing. Fear → relief, pain → parasympathetic rebound, thrill → let-down. The b-process grows with repetition, decays more slowly, and explains post-event “afterfeelings.”
- Pharmacology and addiction. Tolerance plus withdrawal embodies opponent dynamics at their clearest. The direction, latency, and duration of after-reactions follow parameterizable rules.

14.2.3 Neural dampening and gain control

- Habituation and prediction. Limbic responses diminish with repetition; prefrontal and striatal circuits recalibrate value estimates and inhibit overreaction.
- Precision tuning. Neuromodulators (serotonin, noradrenaline, dopamine) regulate gain and precision so that persistent extremes are curbed—an elegant neural basis for adaptation and opponent processes.

14.2.4 Sleep as a normalizer

- REM and emotional memory. Across ordinary weeks, more emotionally intense days are followed by dream-rich nights that reduce next-day reactivity.
- Deprivation effects. Removing REM amplifies limbic responses and slows recovery—precisely the impairment you would expect if sleep normally supports baseline restoration.

14.2.5 Social scaffolding and attachment

- Buffering. Close relationships and reliable support speed return to baseline after stress.
- Economic and cultural modulation. Resources, norms, and status shape how fast the treadmill turns; richer scaffolds compress recovery times.

14.2.6 Developmental smoothing

- Aging and strategy. Older adults often show lower volatility and faster recovery, partly via attentional strategies (positivity effect) and learned regulation—consistent with cumulative tuning of homeostatic and opponent systems.

14.2.7 Quantitative fit with minimal commitments

- Few moving parts. A dual-timescale adaptation term plus an asymmetric opponent captures much of the variance in daily affect and post-event trajectories.
- No global boundary needed. These models do not presuppose any ultimate ledger or terminal neutrality to recover their core effects.

14.2.8 Clear, testable levers

- Manipulability. Sleep, pharmacology, social contact, and cognitive strategies shift rebound speed and amplitude in predictable directions.
- Transfer. Training that reduces reactivity in one domain often generalizes (e.g., from social stress to evaluative threat), as homeostatic control should.

14.2.9 Practical guidance that already helps

- Clinical utility. Behavioral activation, exposure therapy, and sleep hygiene leverage adaptation/opponent dynamics to reduce suffering and stabilize mood.
- Everyday wisdom. “Give it time,” “Sleep on it,” and “Talk to someone” are folk translations of these mechanisms.

14.2.10 Interim verdict in their favor

If one’s evidential horizon is hours to months, under ordinary conditions with intact regulation and support, the hedonic treadmill and opponent dynamics are excellent first-line models. They deliver mechanistic plausibility, quantitative tractability, and practical leverage—without invoking a global fairness constraint. They deserve to be the default against which any stronger claim is measured.

In the next section we turn to the limits: what these accounts cannot guarantee even in their most charitable, modern forms—especially horizon-dependent menu shaping, variance compression near the end of life, systematic dream counterweights tied to prior-day load, and QS-like neural residuals after nuisance modeling.

14.2.11 Where we go next:

Having mapped the strengths, Section 14.3 turns to boundaries: cases where rebound alone leaves residuals—horizon-dependent menu changes, late-life dispersion drops, or cross-stream counterweights—that demand explanation.

14.3 What They Cannot Guarantee

For all their strengths, hedonic adaptation and opponent-process theory have intrinsic limitations. They describe tendencies, not inviolable rules. This distinction becomes clear when we push these theories to their logical and empirical extremes.

First, hedonic adaptation does not guarantee that everyone ends up at neutral cumulative happiness. It's often phrased in probabilistic terms: *most* people *tend* to return toward baseline after significant events. But what about those who don't? Adaptation theory readily admits there are outliers. Long-term studies have identified subsets of people who experience lasting changes in well-being. For example, a substantial subset of individuals show significant, enduring shifts in life satisfaction after major events rather than full reversion. Some people become chronically depressed after traumas and never bounce back to their former baseline; conversely, some individuals find a new lease on life after adversity and stabilize at a *higher* level of well-being. The hedonic treadmill model was refined precisely because evidence showed not everyone stays on the same track – some tracks incline or decline over time. Adaptation theory, being fundamentally statistical, is comfortable with this variability. It can incorporate the idea that set-points differ between people, that genetics and personality yield different “neutral” levels, and that life events can sometimes shift those baselines permanently. In short, it doesn't purport that there's a universal zero-sum ledger across all lives. It says: people have personal equilibria, and they usually return to them, but it doesn't say those equilibria are all zero or that total happiness gained must equal total pain suffered.

What about lifetime balance? Neither adaptation nor opponent theories explicitly requires that, by the end of one's life, all the ups and downs must cancel out. Adaptation focuses on *rates* of return to baseline and magnitudes of residual change; it has no concept of a running “ledger” $L(T)$ that must be settled at death. It is entirely possible under hedonic adaptation to imagine two individuals: one lives a charmed life of mostly positives (with mild adaptation dampening the highs, but still a life that is on the whole positive); another lives a hard life of mostly negatives (with adaptation offering some relief, but still net negative experiences). Adaptation theory does not insist that these two lives must somehow internally balance out to the same cumulative total. It only suggests each will somewhat normalize around *their own* baseline. If one person's baseline mood is unfortunately low due to genetics or environment, adaptation won't magically lift them to neutral; they might spend a lifetime slightly below zero (and adapt to *that* as “normal”). Conversely, a naturally cheerful person in great circumstances might spend a lifetime above zero (though they adapt to good things, they adapt from a higher set-point). In fact,

Diener (2006) explicitly concludes that people are not all hedonically neutral – individuals have different set-points, often non-zero. Thus, hedonic adaptation cannot guarantee *fairness* in the sense of equalized total happiness. It wasn't designed to.

Opponent-process theory, on its own, also falls short of a global guarantee. It says every affective state induces an opposite reaction. But crucially, it does not guarantee that the magnitudes of those reactions are *equal*, especially over long periods or across contexts. Opponent processes are typically modeled as decaying exponential functions – the opponent may not fully cancel out the primary process, especially if stimuli keep coming. For example, while relief follows pain, enough pain can still accumulate trauma that relief cannot erase. Opponent theory doesn't promise that pleasure and pain *integrate* to zero over a lifetime; it only says each instance of one yields some of the other. If someone lives in continuous pain with only brief respites, opponent dynamics might give them moments of relief, but not necessarily enough to offset years of suffering. There's no built-in endpoint where a tally resets to zero. Likewise, if someone experiences repeated high-frequency alternations (e.g. daily stress and nightly relief), opponent-process theory would predict oscillations around a set-point, but not necessarily that the *area under the curve* on each side is equal. The theory is mute about long-term cumulative totals.

Another limitation is lack of horizon sensitivity. Neither adaptation nor opponent mechanisms incorporate the notion of an impending ending or a shrinking future influencing the process. They are *time-scale agnostic* in that sense. Hedonic adaptation doesn't say "as you get older, the speed of adaptation increases" or "when you subconsciously sense the end of life, you adapt differently." It simply reacts to changes whenever they occur. Opponent processes similarly activate in the short-term after a stimulus; they don't look ahead to how many good or bad days you have left. By contrast, LoF specifically posits that horizon H_t – essentially the future time remaining – matters. For example, LoF predicts horizon effects like increased urgency to resolve debts (emotional or interpersonal) as time grows short, and a narrowing of focus to meaningful or comforting activities in end-of-life stages. Standard adaptation theory has no reason to predict such a pattern. In fact, adaptation might suggest the opposite in some cases: an older person might be so well-adjusted to life that they change even *less* in response to events ("I've seen it all; nothing fazes me anymore"), which would not necessarily lead to the specific *rebalancing* push that LoF predicts. Empirically, there are phenomena like the "older adults focus on positive memories" or socioemotional selectivity in aging, but these are usually explained by shifts in motivation and perspective, not by a generic adaptation law. The treadmill by itself doesn't produce a *stronger* counteractive force at the end; it just always produces some counterforce, irrespective of lifespan context.

Perhaps the starker illustration of what these theories cannot guarantee is to imagine a thought experiment: Suppose we have a perfectly “open-channel” scenario (no external barriers to emotional experience or expression) and we let someone’s life run its course. Could they end up with a large positive surplus of happiness, or a large negative deficit of happiness? Adaptation and opponent processes would *tend* to nudge them toward equilibrium along the way, but they wouldn’t strictly prevent an imbalance. It’s conceivable, under these theories, for someone to die peacefully in old age having had, say, 60% positive experiences and 40% negative – a net positive life. Adaptation would say they likely got used to the good fortune and maybe didn’t feel as ecstatic by the end, but it doesn’t retroactively create suffering to balance it out. Similarly, someone could, in theory, end their life with more pain than pleasure accumulated; adaptation might have kept them from continuous misery (they adapt to each hardship), but still, the ledger could be negative if hardships kept coming. LoF, in contrast, boldly asserts that (in the absence of impediments) such net imbalances *won’t persist* by life’s end – something will happen to drive the total toward neutral. That is a claim of *guarantee*, not just tendency.

One more point they cannot guarantee: cross-domain compensation. Hedonic adaptation tends to be domain-specific – you adapt to your new car, your marriage, your health condition, largely within that domain of experience. Opponent processes are typically modality-specific – the rebound is often directly related to the initial stimulus (fear to relief, drug high to withdrawal). But LoF, as proposed, suggests something broader: if one channel of relief is blocked, another might take over; dreams might counterbalance waking life; a person might find closure in a completely different arena than where the pain came from (e.g., someone with chronic illness finds emotional balance through art or relationships). Standard adaptation theory doesn’t guarantee that *if one pathway fails*, another will compensate – it would just observe that failing to adapt in one domain means the person stays unhappy. LoF would imply the system *finds a way* to balance, perhaps through less direct means, provided consciousness continues and some “channels” remain open (be they social, cognitive, creative, etc.). This kind of orchestrated compensation across domains is beyond the remit of adaptation or opponent theories as traditionally formulated.

In summary, hedonic treadmill and opponent-process theories cannot guarantee the strong invariances that LoF proposes. They cannot promise that every life, when given every opportunity, ends in a hedonic draw. They cannot promise invariant patterns tied to the approach of death (like a mandatory narrowing of variance or final acts of closure). They allow for exceptions, they lack a built-in horizon trigger, and they focus on local dynamics rather than global outcomes. This doesn’t make them “wrong” – it just means

they are different in character from a law. They're more like weather patterns than a climate law: useful for predicting common fluctuations, but not asserting an exact long-term balance.

LoF explicitly goes beyond by saying *if the system is working properly*, certain outcomes *must* occur (within defined margins of error)—for instance, the final ledger mean would fall within a preregistered equivalence band around zero with compressed variance, subject to measurement invariance and covariate control. Adaptation theory would never make such a specific claim; it's content with saying “likely around baseline, but who knows how close.” This gap — between “usually” and “necessarily” — is precisely where LoF claims its territory.

14.3.1 Horizon-dependent menu shaping (the admissible set)

- The gap. Adaptation and opponent models describe what happens after you feel something, not which options are present before you act. They modulate reactivity, not the availability and stickiness of options.
- LoF requirement. As personal horizon H_t shrinks (e.g., late life, impending closure), the *menu itself* is predicted to narrow and tilt toward reparative, ledger-improving actions. In LoF, this is the Queue System’s admissible-set dynamics; in rivals, there is no principled way for horizon to filter thoughts while preserving ordinary agency.
- Discriminator. Equate immediate utility, conflict, arousal, and risk; vary only horizon framing. If choice sets and neural commit thresholds still tilt toward repair as H_t shortens (vmPFC value boost for reparative options; rIFG/ACC brakes on indulgent or debt-deepening options), that is beyond classic adaptation. Treadmill models predict faster rebounds after the fact, not pre-emptive menu curation.

14.3.2 Terminal variance compression and neutral closure

- The gap. Treadmill/opponent accounts imply tendencies toward baseline, but say nothing about end-of-life dispersion of cumulative affect. They allow wide lifetime ledgers provided local rebounds occur.
- LoF requirement. As death of mind approaches, cumulative affect is predicted to show variance compression toward a preregistered neutral band, with compensatory intensification (e.g., reconciliation urgency, peace phenomena, last-acts concentration) detectable in HCl trajectories, language, and autonomic indices.

- Discriminator. If well-measured streams exhibit population-level narrowing of net affect as biological closure nears—beyond what illness-severity, sedation, and social factors explain—rival theories lack parameters that force such global convergence. They can fit some cases post hoc but not guarantee it in principle.

14.3.3 Ledger-tuned dream counterweights

- The gap. Opponent dynamics can explain after-swing in physiology and mood, and sleep research supports general emotional normalization. But they do not mandate content-specific counterweights that track the signed load of the prior day or week.
- LoF requirement. Dreams function as low-cost counterweights: after unusually negative days, we should observe elevated positive dream valence or successful mastery themes; after unusually positive days, proportionally more threat rehearsal, loss, or humility themes—with magnitudes scaling to ledger imbalance and horizon.
- Discriminator. In high-compliance REM awakenings with blind coding, show a monotone, dose-responsive inversion of dream valence against prior-day HCl, independent of circadian stage, substance use, and expectancy. Rival models predict some normalization, not a calibrated, ledger-sensitive inversion.

14.3.4 QS-like residuals in control hubs after nuisance modeling

- The gap. Adaptation accounts can explain limbic dampening and prefrontal inhibition as generic control. They cannot explain a residual valuation/inhibition term that tracks feasibility-of-compensation (Φ) after we regress out utility, conflict, arousal, habit strength, and risk.
- LoF requirement. vmPFC/OFC should show value boosts for options that improve neutral-closure probability; rIFG/ACC should raise commit thresholds for options that worsen it—specifically as a function of $\Phi \times$ Horizon.
- Discriminator. If preregistered models reveal a stable Φ -residual in vmPFC and horizon-interaction in rIFG/ACC across tasks and individuals, rivals need to add a new state variable tantamount to a fairness shadow price, which is precisely the LoF mechanism.

14.3.5 Population dispersion bounds and cross-stream coupling

- The gap. Treadmill/opponent theories are within-stream: they describe how one person's affect drifts. They do not supply population-level coupling ensuring that

no stream accumulates unbounded hedonic surplus or deficit by the end of its conscious span.

- LoF requirement. There must exist dispersion bounds on final ledgers across the population, and cross-stream constraints when streams interact: my admissible set can tighten if your choices would otherwise force me into uncompensable debt, and vice versa.
- Discriminator. In multi-agent telemetry, observe coordinated tilting of menus under conflict—repair options for A gaining stickiness precisely when B’s horizon shortens or ledger worsens—without explicit communication or shared incentives. Rival models can fit strategic cooperation; they do not predict latent compensation coupling driven by a fairness constraint.

14.3.6 Summary table — rival predictions vs. LoF requirements

Domain	Rival accounts predict	LoF requires	Empirical discriminator
Menu shaping	Post-hoc rebound; unchanged pre-choice menu	Horizon-dependent admissible set with reparative tilt	vmPFC value boost and rIFG/ACC brakes vary with $\Phi \times H$ holding utility/conflict constant
End-of-life	No constraint on lifetime dispersion	Terminal variance compression toward zero ledger	HCI narrowing near biological closure beyond clinical covariates
Dreams	General normalization	Ledger-tuned valence inversion	Blind-coded REM reports track prior-day HCI with dose-response
Neural residuals	Generic inhibition/habituation	Φ -specific residual after nuisance regressors	Stable QS-residual across tasks, preregistered
Cross-stream	Independent drifts; social effects only	Coupled admissible sets to avoid uncompensable debt	Coordinated repair tilts in interacting dyads/groups

14.3.7 What would count as decisive for the rivals

If, after rigorous nuisance modeling and preregistered tests:

1. No Φ -residuals are found in vmPFC/ACC/rIFG;
2. No horizon interaction appears in menu shape or commit thresholds;
3. No terminal compression is seen in well-measured streams;
4. Dream valence shows normalization without ledger-sensitive inversion; then a well-tuned adaptation/opponent framework would be sufficient. In that case LoF reduces to a useful metaphor rather than a law.

14.3.8 What would force the rivals to import LoF

Conversely, if those signatures do appear robustly, rival accounts must add a global state variable that shapes option availability as a function of compensability and horizon, and must impose boundary conditions on path space that compress terminal dispersion. At that point they cease to be true alternatives and become LoF in disguise.

Bottom line: adaptation and opponent processes excel at explaining rebounds within life, but they do not and cannot by themselves guarantee the end-of-life neutrality, horizon-dependent menu shaping, ledger-tuned dreams, QS-like neural residuals, or population dispersion bounds that define the Law of Fairness. Those demands require a constraint law—or else the data will tell us, cleanly, that no such law exists.

14.3.9 Where we go next:

To separate “tendency” from “law,” Section 14.4 lays out preregistered tests and invariance checks. We specify thresholds, negative controls, and one out-of-sample metric to keep the goalposts fixed.

14.4 Research Notes: Tendency vs. Law

The distinction between a tendency and a law can be sharpened with a bit of formal thinking. A tendency (like hedonic adaptation) is a *statistical regularity*: one might say $P(\text{returning to baseline} \mid \text{enough time})$ is high, or that the expected deviation decays over time. A law, in the sense of LoF, implies an *invariant outcome* under specified conditions (open channels, intact cognition, etc.). We can articulate this difference through the concept of the life ledger $L(T)$ defined as the time-integral of net affect up to time T (i.e. cumulative pleasantness minus unpleasantness experienced). LoF essentially posits that for a conscious life that reaches its natural terminus *with opportunities for compensation intact*, $L(T)$ approaches zero (neutral) as T approaches the end of life. Adaptation, by contrast, would only say that $L(T)$ is likely to grow sublinearly or level off, since positive and negative deviations tend to be dampened. But it *does not insist* on zero.

To make this concrete: imagine measuring a person's well-being on a standardized composite scale (say the Hedonic Composite Index, HCI) continuously over their life. Adaptation hypothesis: this person's HCI will tend to fluctuate around some set-point and, after shocks, will revert towards that set-point. Over a long time, the distribution of their HCI might be roughly stationary or at least bounded. LoF hypothesis: if we integrate that HCI curve from birth to the end of life, the total area above zero will approximately equal the total area below zero (assuming we weight it so that neutral registers as zero net). LoF further implies that as the end approaches, the system predictably shifts affective trajectories toward neutral within preregistered bounds, conditional on intact cognition and open compensatory channels – for instance, by last-minute surges of meaning-making or emotional equilibration. This is a much stronger condition.

From a testing standpoint, one can formulate competing models and compare out-of-sample fit. For example, consider a logistic regression for the probability of choosing a “repair” or compensatory action (e.g., making amends, seeking relief) at time t . LoF predicts that this probability depends not just on the immediate context, but on an interaction between perceived remaining horizon H_t and compensability Φ_t (feasibility-of-compensation: how much an action can repair, relieve, or preserve reversibility relative to the current ledger state). Symbolically, we can write:

$$\text{logitP(choose repair/relief at } t\text{)} = \alpha + \beta H g(H_t) + \beta \Phi \Phi_t + \beta \{g(H) \times \Phi\} (g(H_{-t}) \times \Phi_{-t})$$

where $g(H_t)$ is a preregistered monotone “urgency” transform that increases as the horizon shrinks (e.g., $g(H_t) = H_t^{-1}$, defined with safeguards to avoid division-by-zero). Under a pure adaptation or homeostatic account, Φ_t may influence choices and horizon

may have independent effects, but there is no requirement that shrinking horizon amplifies the pull of high- Φ actions. By contrast, LoF predicts $\beta_{\{H \times \Phi\}} > 0$ for the interaction term $g(H_t) \times \Phi_t$: as the horizon shrinks (urgency rises), the weighting on high- Φ options increases. Finding a reliable $g(H_t) \times \Phi_t$ interaction in well-powered, preregistered studies would support LoF beyond ordinary adaptation; failing to find it under clean conditions would favor the rivals.

such a model, a pure adaptation or homeostatic theory does not require any special interaction term $\beta H : \Phi$ – an action’s value might influence choices, and horizon might have independent effects (e.g., maybe older people are more patient, etc.), but there’s no necessity for *horizon to amplify the drive toward high- Φ (compensatory) actions*. LoF, by contrast, *does* predict a specific interaction: as the horizon H shrinks, the weighting on high- Φ actions should increase (i.e. $\beta H : \Phi > 0$ if Φ represents something like “this action helps settle the ledger”). Finding a significant $H \times \Phi$ term in well-powered studies would support LoF’s law-like effect beyond ordinary adaptation. Not finding it (when one should, under clean conditions) would favor the rivals.

Another empirical discriminator: menu size and variance. Adaptation says people return to feeling normal, but doesn’t necessarily say their *behavioral repertoire* narrows as they near life’s end. LoF predicts that as time grows short, not only does mood center around neutral, but people’s set of actions (the “menu”) *shrinks* to mostly essential, restorative ones. One can test this by counting the variety of activities or goals pursued by individuals with long vs. short horizons (e.g. younger adults vs. terminally ill older adults with full mental clarity). If we see no systematic reduction in the variety of activities when controlling for energy and context – in other words, if people continue to explore and engage in a wide range of behaviors even very late in life – that would align with adaptation (no law forcing narrowing) and conflict with LoF’s prediction. Technically, one could model activity counts with a Poisson or Negative Binomial model and check for an effect of horizon. LoF expects a menu-tightening effect: significantly fewer distinct pursuits as H decreases, beyond what fatigue or disability alone would cause. Adaptation alone might predict some emotional re-centering but “business as usual” in decision-making breadth. If data showed, for instance, that a healthy person who believes they have a short time left drastically limits their goals to just a few meaningful ones (whereas someone with a long open horizon keeps a longer list of diverse goals), that supports LoF’s unique claim.

Fail pattern (No Menu Narrowing): If we observed that as people approach the end of life (with awareness of it), they *do not* streamline their activities or priorities – e.g. they continue to seek novel experiences and chase long-term ambitions just as much as

before – it would undermine LoF's horizon effect while being quite compatible with standard adaptation (which imposes no such constraint). In formal terms, if a well-designed study found that the number of distinct daily activities or goals remains constant (or even increases) as $H \rightarrow 0$ after accounting for health and fatigue, then LoF's prediction of horizon-driven menu tightening fails. Adaptation would shrug: it never predicted any tightening in the first place.

We can also examine end-of-life affective profiles. LoF delineates a very specific profile for the final phase (when “channels” for emotional processing are open, e.g. hospice with good psychosocial support): the person’s affective state should converge to near-neutral (within some small margin, like ± 0.15 SD) with a *flattening of slope* (no systematic upward or downward drift) and a *compressed variance* (maybe $\leq 80\%$ of mid-life variance). Adaptation, by contrast, does not entail any particular behavior of variance or mean near death. An adaptation theorist might expect that if someone’s circumstances at end-of-life are stable and they’ve adapted to them, their mood is near their personal baseline – but that baseline could be above or below neutral, and the variability could be anything (some people might be highly volatile if they have pain bursts, others might be flat). Adaptation certainly doesn’t guarantee the *absence* of despair or euphoria near death; it would only say those extreme feelings might dampen over time. So a key empirical test is: do we see a reliable pattern of *neutral, calm emotional states* in people who are dying but mentally lucid and not acutely suffering physically (i.e., the situation where LoF should operate)? If yes, and particularly if the data show a trend toward neutrality as the time horizon shortens in each individual, that’s evidence in favor of LoF. If not – if many people end their lives with persistent high happiness or deep misery despite having support and awareness – that challenges LoF. Adaptation alone would not be falsified by either outcome, since it has no strict expectation there.

Fail pattern (No End-of-Life Convergence): Suppose rigorous measurements showed that even under excellent care, about half of individuals maintained a strong positive or negative bias in their mood right up to the end – say some remained elated and some remained despondent, with no overall compression toward neutrality. Statistically, if the final weeks’ average mood fell outside the ± 0.15 z equivalence band around zero, or if variance in affect remained just as wide as earlier in life, this would count against LoF but not against adaptation. Hedonic adaptation would explain it as individual differences in set-points or coping, whereas LoF would struggle because it predicts a consistent convergence that just didn’t show up. Such an outcome (no convergence) is a scenario where the rivals “win”: the balancing observed earlier in life would appear to not be a universal trajectory but just a general tendency that can fail in the endgame.

Finally, a model comparison approach can quantitatively assess tendency vs. law. We can pit a model embodying adaptation/opponent processes against a model including the LoF constraint, and then evaluate which predicts new data better. Modern statistical tools like cross-validated log-likelihood, WAIC, or LOO can compare the out-of-sample performance of these models. If the adaptation-based model explains all the variance and the LoF-augmented model offers no improvement (or is beaten in predictive accuracy), then adding LoF is unnecessary. Conversely, if the LoF-informed model yields systematically better predictions (especially on those signature end-of-life or horizon-interaction patterns) without overfitting, that's evidence for the law-like effect. The key here is *preregistered* tests on fresh data – to avoid bias, one would specify in advance the criteria like “we expect final-week well-being mean to be within X of neutral and variance lower than Y” and then see if reality matches it. The research notes in LoF emphasize exactly this kind of adversarial collaboration: fit the rivals first, let them predict what happens, and see if LoF's extra term (like that $H \times \Phi$ interaction) truly adds explanatory power.

To sum up, the tendency vs. law distinction is not just semantics – it yields different research strategies. A tendency (adaptation) is something we observe broadly but with tolerance for exceptions; a law (LoF) demands we look for invariants and be ready to declare the law falsified if those invariants fail consistently. Thus, this chapter's research notes outline how one would *falsify LoF specifically*: e.g., find that no horizon-dependent effect actually occurs, or that some people's ledgers remain unbalanced to the end with no system correction. It also outlines what would falsify the idea that “it's just adaptation”: e.g., find robust evidence of horizon \times imbalance interactions and end-of-life equivalence profiles that adaptation models did not predict. By specifying these outcomes in advance, we ensure the competition between tendency and law is decided by data, not by rhetorical preference.

Ethical Note: In all such tests, the guiding principle is that *relief is a systems variable, but a person's comfort and dignity override data collection*. We will never, for instance, withhold analgesics or emotional support to “see what happens” if suffering goes unchecked – that would be grossly unethical and scientifically meaningless. All observations must be made with minimal intrusion and maximum compassion. In practice, this means end-of-life studies rely on *observational* or *retrospective* data (or gentle, volunteer-driven tasks), and any attempt to explore these dynamics experimentally must prioritize participants' well-being above any hypothesis. The Law of Fairness, if real, should reveal itself in patterns we can detect without ever denying someone comfort.

14.4.1 Competing hypotheses (formal)

- Hypothesis (T): Local affect follows autoregressive homeostasis with opponent after-reactions. Choice sets are unaffected by global ledger or horizon except via ordinary fatigue, risk, and utility. No constraint on lifetime dispersion. Minimal model:
- $F_{t+1} = \alpha F_t + \sum_k \theta_k X_{k,t} + \eta_t$, with $|\alpha| < 1$, opponent terms in X , and no global boundary term.
- H_law (L): Affect dynamics include a constraint term tied to ledger $L(t)$ and horizon H_t that (i) filters options (admissible set $\mathcal{A}(t)$), (ii) induces horizon \times compensability interactions in valuation/inhibition, and (iii) produces terminal variance compression of $L(T)$ about zero. Minimal augmentation:

$F_{t+1} = \alpha F_t + \sum_k \theta_k X_{k,t} + \lambda_t \Phi_t + \eta_t$, $\lambda_t = g(H_t)$, $\Phi_t = \text{compensability}(u_t; L(t), H_t)$, plus a population boundary: $\text{Var}[L(T)] \rightarrow c$ within a preregistered bound under adequate measurement.

14.4.2 Five law-level signatures (with effect metrics)

1. Horizon-dependent menu shaping
 - Metric: change in choice-set entropy and repair-option weight as a function of H^{-1} , controlling utility/conflict/arousal.
 - Report: $\Delta \text{Entropy}(\mathcal{A})/\Delta(H^{-1})$; vmPFC value residual $\beta\Phi \times H$; rIFG/ACC commit-threshold $\gamma\Phi \times H$.
2. Terminal variance compression
 - Metric: slope of population variance in cumulative HCl versus proximity to biological closure (days–weeks).
 - Report: $(d/dt)\text{Var}[L(t)] < 0$ near end, after covariate adjustment (illness severity, sedation, communication access). Target partial $R^2 \geq .05$ with cross-site replication.
3. Ledger-tuned dream inversion
 - Metric: regression of REM dream valence index on prior-day HCl with negative slope that scales with $|L(t)|$ and H^{-1} .
 - Report: $\beta_{\text{dream} \leftarrow \text{HCl}_{t-1}} < 0$, interaction $\beta \times |L|$, preregistered REM windows.
4. QS-residuals after nuisance modeling

- Metric: vmPFC/OFC value and rIFG/ACC inhibition explained by Φ after controlling utility, conflict, risk, habit, surprise, and effort.
- Report: added-explained-variance $\Delta R^2 \Phi \geq 1\text{--}2\%$ in ROI GLMs, with held-out prediction.

5. Cross-stream coupling

- Metric: dyad/group menu tilt synchrony when ledgers/horizons diverge, independent of explicit incentives.
- Report: cross-lagged path coefficients A→B and B→A for repair-option weight, conditioned on communication/content.

14.4.3 Statistical posture: “law” requires more than significance

- Evidence thresholding. For each signature, require replicable effects across ≥ 3 labs, pre-registered, with $BF_{10} > 30$ (strong) or meta-analytic Cohen’s d or partial R^2 with heterogeneity $I^2 < 40\%$.
- Convergence across modalities. Each signature should appear in behavior, neural (EEG/fMRI), and passive telemetry. LoF is not a single-task effect.
- Adversarial collaboration. Design and analysis are co-owned by LoF and rival teams; joint preregistration; symmetric veto on analytic flexibility.

14.4.4 Model comparison and “tendency penalty”

- Baseline models: ARIMA/State-space with adaptation + opponent + risk + utility + habit + circadian + sleep.
- LoF model: Baseline plus Φ , H, and their interaction; admissible-set filter on options; terminal boundary on cumulative dispersion.
- Information criteria: Prefer WAIC/LOO-CV and out-of-sample predictive log-likelihood.
- Tendency penalty: If LoF only yields in-sample fit and not out-of-sample gains, it fails as a law (too many knobs).
- Ablations: Remove Φ , remove H, remove admissible filter; quantify degradation $\Delta WAIC$ and Δ predictive R^2 .

14.4.5 Optional stopping and martingale framing

- To avoid spurious “compression,” use fixed sample sizes or alpha-spending.

- For terminal neutrality, use optional-stopping-robust analyses (e.g., nonparametric confidence sequences) and make explicit the required regularity conditions for any martingale-based guarantees.

14.4.6 Measurement discipline

- Invariance audits. For HCI, demonstrate configural → metric → scalar invariance across cohorts and time. If scalar fails, use partial invariance with alignment optimization; propagate uncertainty into $L(t)$.
- Blind coding and dual pipelines. Independent teams score dream content, end-of-life language, and repair themes; reconcile only after lock.
- Negative controls. Include variables predicted to not correlate with Φ (e.g., color-word Stroop accuracy when utility/risk equalized).

14.4.7 Minimal effect sizes (pre-specify)

- Menu entropy shrinkage near closure: $\geq 10\text{--}15\%$ reduction per SD of H^{-1} after covariate control.
- vmPFC Φ -residual: added $R^2 \geq 1\%$ is realistic and meaningful in fMRI; reproducible across tasks.
- Dream inversion slope: standardized $\beta \leq -0.10$ with REM-only windows and blind coders.
- Terminal variance compression: slope $< -0.02 \text{ SD}^2/\text{day}$ over the last ≤ 14 days in hospice datasets, after covariate adjustment.

14.4.8 Robustness grid

- Structural: swap Φ features (relief, repair, risk, reversibility); effect persists.
- Parametric: priors widened/narrowed; leave-one-site-out; effect persists.
- Algorithmic: GLM vs. GAM vs. hierarchical state-space; effect persists.
- Implementation: different ROIs/parcellations; different dream coding schemes; effect persists.
- Epistemic: analysts blinded to LoF vs. rival labels; conclusions stable.

14.4.9 Identifiability and DAGs

- Causal DAG includes: Utility → Choice; Conflict → Inhibition; Ledger and Horizon → Φ → Admissible Set → Choice; plus Sleep/Context confounds.
- Use front-door/IV strategies where possible (e.g., exogenous horizon cues) to identify the Φ pathway distinct from utility/risk.

14.4.10 Decision rule for reviewers

- Support LoF as a law if ≥ 3 signatures meet thresholds with adversarial, preregistered, cross-lab replication, and if rivals cannot match predictive performance without importing a functionally identical $\Phi \times H$ term and terminal boundary.
- Downgrade to tendency if effects are small, fragile, or vanish under invariance corrections and nuisance modeling, or if rivals achieve equal prediction with homeostatic parameters alone.
- Reject if negative controls light up, if compression flips sign when covariates tighten, or if preregistered cross-validated metrics favor rivals.

14.4.11 Practical templates (ready to copy into preregistrations)

- Primary endpoint: cross-validated Δ log-likelihood for LoF vs. baseline in predicting repair-option selection under horizon framing.
- Key moderator: H^{-1} from time-to-deadline or actuarial prognosis.
- Neural ROIs: vmPFC/OFC for value residual; rIFG/ACC for commit/inhibit; insula for interoceptive “rightness.”
- Dream protocol: 7-night REM awakenings, blind coders, a priori lexicons for agency, mastery, care, threat, humility; analysis by multilevel models with person-level random slopes.

14.4.12 Where we go next:

If the rivals pass those tests, what follows? Section 14.5 sketches what LoF should learn or surrender—methods to keep, claims to retire, and a clean handoff to standard models where they suffice.

14.5 If Rivals Win, What LoF Learns

What would it look like for the hedonic treadmill and opponent processes to “win” against the Law of Fairness? It’s worth painting that scenario, because it frames how we approach this whole inquiry. If, after rigorous testing, we find that all the characteristic signatures we thought might evidence LoF are actually explicable by these established mechanisms, then the conclusion would be that LoF as a separate law *is not needed*. In scientific terms, we’d favor the *parsimony* of existing theory: why posit a new law if the old ones suffice? Concretely, imagine we conduct a battery of studies and analyses over the coming years. We measure horizon effects in decision-making, emotional trajectories in hospice patients, dream contents after stressful days, neural signals of “unresolved” experience – the whole gamut. Now suppose the results consistently show that no additional LoF factor is required: every pattern falls out of known processes. For example, perhaps older or terminally ill individuals do show some narrowing of activities, but it turns out to be fully explainable by changing priorities or energy (which socioemotional selectivity theory or basic adaptation could cover) rather than a fairness drive. Or suppose dreams after hard days sometimes lighten mood, but we find this is just because REM sleep naturally modulates emotion and memory, not because it specifically “compensates” in a ledger sense. And let’s say our model comparisons favor the rivals – e.g., an adaptation-based model with a bit of randomness predicts emotional data as well as any LoF-augmented model, yielding no improvement in WAIC or cross-validated log-loss when the “fairness constraint” is added. Furthermore, let’s assume we observe some clear counter-examples to LoF: perhaps a subset of people (with no obvious differences in available support) end life on a markedly positive or negative note without converging, and this can be traced to idiosyncratic but non-mysterious causes (like stable personality traits or social factors). In short, *suppose the rivals can replicate all the LoF signature phenomena through their own terms, and any unique prediction of LoF fails to materialize.*

In that scenario, LoF would effectively collapse into the broader umbrella of existing theory. We would conclude that what we’ve been calling a “law” is really an emergent consequence of multiple adaptive processes working in concert – or perhaps even an illusion created by selection bias in what cases we’ve observed. The “tendency toward fairness” would remain as a descriptive phrase, but we wouldn’t elevate it to fundamental status. Importantly, this outcome is *not* a loss for science or humanity. On the contrary, it’s a clarifying win. It would mean we have mapped out the limits of hedonic adaptation and homeostasis more clearly. We would have a better understanding of where those processes suffice and where they don’t. Maybe we’d learn, for instance, that *given sufficient time*, adaptation *almost always* equalizes things, but if life is cut short

suddenly, there's no mysterious force to guarantee balance – it was just time and normal coping all along. Or we might learn that people with rich social support networks achieve something close to LoF's neutrality at end-of-life, but that's because social and psychological factors provided channels for every last adjustment (again, nothing beyond known factors).

If rivals win, we also learn something about human resilience and limitations. It could imply that our natural adaptive capacity is even more impressive than we thought – that even the end-of-life reconciliation and dream counterweight patterns can be generated by the same mechanisms that help us recover from everyday setbacks. That would be a profound insight: it would mean nature equipped us with all we need in terms of emotional homeostasis, without requiring an extra principle. It might also direct our efforts toward bolstering those natural mechanisms (since they do the job when conditions allow). For instance, if social baseline theory fully explains end-of-life peace (perhaps just having companionship and reduced stress is enough to account for “ledger closure”), then interventions would focus on ensuring everyone has that social support and stress reduction, rather than chasing a new fundamental balance principle.

By stress-testing adaptation and opponent-process theories at their extremes, we either affirm their sufficiency or find their breaking points. If they hold up, we return to the theoretical table with a clearer understanding: hedonic adaptation is robust, but maybe we discover exactly under what conditions it fails or succeeds. Opponent-process theory might gain new domains of validation (e.g., perhaps dreams do show an opponent-process effect after all, in purely neural terms). We likely would have developed better measurement tools (like the HCl metrics, longitudinal methods) in the process of trying to test LoF, and those remain useful.

It's also worth noting that “rivals win” doesn't necessarily mean every pattern is explained by one single rival theory. It could be that some combination of them – say, adaptation + predictive coding + social support dynamics – together account for everything. In that case, LoF's lesson would be that it was pointing toward a *cluster* of processes already known, and the apparent simplicity of a single law might dissolve into a more complex, but still understandable, set of factors. The end result might be an integrated model of emotional life with no mysterious remainder. That is still a win for knowledge.

So how would we know the rivals have truly won? We would have pre-specified criteria for LoF that fail. In Part VII and beyond, we've set up what those criteria are: for example, no significant horizon-by-compensation interaction in well-powered tests, no evidence of systematic end-of-life neutrality beyond what adaptation predicts, and perhaps a

head-to-head model comparison where an LoF-informed model does not outperform a well-tuned adaptation/homeostasis model. If those come to pass, the fair-minded conclusion is that LoF as a separate principle is unnecessary. We would say, in effect, “We tried to break the existing theories by looking at the toughest cases; they held up. There’s nothing extra guiding the scales – just the known processes doing their thing.”

The tone of this outcome is not one of tragedy or defeat. It’s actually somewhat reassuring: it would mean that the intuitive sense of “people bounce back” was essentially correct and complete. It would also mean there’s no hidden cosmic bookkeeping of joy and suffering – life is just what it is, and any fairness we perceive emerges from mundane mechanisms. Ethically, it might prompt us to focus even more on ensuring those mechanisms (like social connection, coping resources) are available to everyone, because there’s no automatic guarantee of balance if they’re absent. In fact, one could argue that if LoF fails, *human intervention* and care become all the more important – because we can’t rely on a built-in law of nature to make things right in the end.

In the spirit of this project’s honesty, we commit to acknowledging if the rivals win. The aim was never to score a victory for a pet theory; the aim was to understand the truth of our emotional universe. As the “Bottom line” in our technical notes put it: “*Adaptation is the engine; LoF claims a guardrail. If all we ever see are returns to baseline, rivals win. If we see horizon-specific tilts, menu tightening, and end-game compression that rivals can’t match, then adaptation alone isn’t the whole story. That difference is empirical — and that is how we will settle it.*” In Part VII, we have set the stage for that empirical test. If the outcome is that the guardrail was imaginary and the engine was sufficient, then LoF gracefully bows out as a standalone idea. We still learn *why* life often feels fair in the small sense (because of adaptation and our psychological immune system), and we learn the limits of that fairness (cases where it fails, which might inspire compassion and policy changes).

14.5.1 What “rivals win” would actually mean

“Rivals win” is not one vague outcome; it is a pattern of results:

1. Menu-invariance: Choice-set structure shows no reliable dependence on horizon beyond ordinary variables (utility, conflict, fatigue, risk-aversion). The hypothesized admissible-set tilt with H^{-1} is absent or explained entirely by standard control-theory terms (e.g., opportunity cost, risk sensitivity).
2. No terminal compression: Cumulative hedonic dispersion does not shrink as biological closure approaches when measurement invariance is enforced and

clinical covariates are modeled. Observed “peacefulness” near death reduces to analgesia, sedation, and selection effects.

3. No QS residuals: After nuisance modeling, Φ -based regressors add no predictive value in vmPFC/OFC (value), rIFG/ACC (control), or behavior. Any apparent residuals dissolve under stricter cross-validation, negative controls, and multi-site replication.
4. Dream neutrality: REM content valence does not invert as a function of ledger or horizon once sleep architecture, circadian effects, and stress hormones are controlled. Counterexamples dominate.
5. Rival sufficiency: Adaptation + opponent processes + predictive-coding/FEP + standard RL explain all task-level and telemetry phenomena with fewer parameters or better generalization.

If this pattern holds across labs and pre-registrations, LoF as a law does not meet its own preregistered criteria for law-like status. What, precisely, would we learn?

14.5.2 Lessons about affect and control (if rivals prevail)

- Local mechanisms suffice. Homeostatic control and efficient coding may fully account for affective rebounds and apparent “balancing,” without any global constraint that targets lifetime neutrality. The nervous system economizes prediction error and metabolic cost; hedonic set-points may drift but do not obey a global ledger.
- Horizon is ordinary cost, not a shadow price. Time-to-goal effects likely reflect standard discounting, opportunity cost, and hazard estimation, not a fairness-regulatory multiplier. The “horizon effect” then belongs to canonical decision science, not to a fairness constraint.
- Counterweights are optional, not obligatory. Dreams, rituals, and social reconciliations may serve many functions (memory consolidation, social grooming, norm display) rather than enforcing a compensatory law. Their presence or absence would vary with ecology, culture, and physiology.
- Terminal affect is contingent, not conserved. End-of-life trajectories might be dominated by medication, oxygenation, delirium, attachment dynamics, and cultural scripts. “Peace” is then an achievement of care, not a consequence of a hidden boundary condition.

14.5.3 What LoF should revise or relinquish

- Relinquish the neutrality guarantee. If terminal compression and admissible-set tilt fail decisively, the claim that *every* conscious stream resolves to approximate hedonic zero at death must be withdrawn. LoF would no longer be a law; at best it would become an interpretive ethos or a normative aspiration.
- Retire the Queue System as a regulator. If no $\Phi \times H$ residuals remain after robust controls, the QS hypothesis (a constraint that prunes the pre-choice menu to protect lifetime neutrality) should be abandoned as a mechanism. We would revert to standard control-hub accounts (ACC/rIFG) with no special fairness role.
- Narrow the domain of the ledger. The Life Ledger becomes an analytic summary, not a conserved quantity. It may still be useful for narrative and clinical insight, but not as a target of lawful regulation.
- Demote simulations to educational tools. If adversarial model suites show no distinctive LoF signatures that survive stress-tests, the simulation program should be reframed as a way to teach decision science and ethics, not to discover a new law.

14.5.4 What is still worth keeping (even if LoF loses)

- Measurement craft. The Hedonic Composite Index (multi-modal affect measurement with invariance audits, blind pipelines, and preregistration) remains a field advance. Rivals also benefit from better meters. Keeping rigorous configural → metric → scalar checks is a net gain.
- Ethical guardrails. The strict stance on euthanasia paradigms, coercion, sensitive telemetry, and posthumous dignity does not depend on LoF being true. These protections should remain standard for any serious work on consciousness and suffering.
- Catalog of edge cases. The unified framework for identity (unity-by-access, stream adjudication in split-brain, DID, coma, organoids, and AI) is independently valuable for philosophy of mind and clinical triage.
- Hypothesis-generating metaphors. The “menu” and “ledger” metaphors can enrich pedagogy in psychiatry and life-course counseling, even if they are not literally conserved or regulated by a law.

14.5.5 What research the field should do next (if rivals win)

- Strengthen rival mechanistic bridges. Push adaptation/opponent models to the mesoscale: derive predictions from neuromodulator dynamics (DA/5-HT/NE/ACh), metabolic constraints, and predictive-coding hierarchies for specific tasks and clinical states.
- Clarify dream functions mechanistically. Double-down on memory-consolidation and threat-simulation accounts with better REM sampling, endocrine panels, and manipulation (e.g., prazosin for nightmares, targeted memory reactivation).
- End-of-life science with humility. Build richer causal models for terminal affect (pain control, hypoxia, dyspnea, microbiome shifts, social presence) and test interventions that raise dignity without invoking fairness constraints.
- Map control hubs without QS. Use TMS/TUS/DBS, lesion studies, and computational psychiatry to model commit/inhibit dynamics as outcomes of conflict monitoring, expected value of control, and risk management—no global ledger needed.

14.5.6 Philosophy: reframing fairness without lawhood

If rivals win, “fairness” should be reinterpreted as a normative project rather than a nomic principle. We can still argue for institutions and practices that approximate hedonic justice—palliative care, trauma-informed policy, equitable access to analgesia, social repair—but we must stop claiming that the world guarantees it. The moral force then arises from human commitment, not from metaphysical necessity.

14.5.7 What would change in this book

- Labels: Chapters describing QS as a regulator would be rewritten as test proposals and negative results, preserving the methods but removing claims of lawhood.
- Conclusions: The final statement would shift from “fairness as a fundamental law of experience” to “fairness as a compelling research program and societal aim,” with clear boundaries.
- Appendices: All preregistrations, null results, and adversarial analyses would be published in full, strengthening credibility and enabling rivals to build on the measurement platform.

14.5.8 One page, one promise

If rivals win, we promise three things. First, we will publish every null, every failed replication, and every analysis that undercuts LoF. Second, we will keep and improve what worked—composite meters, invariance audits, ethical scaffolding—and donate these tools to the field. Third, we will say plainly what the evidence then says: that the human nervous system tends toward regulation through familiar mechanisms, that any appearance of lifetime balance is contingent, and that the burden of fairness rests with us.

In conclusion, a “rivals win” scenario teaches LoF (and us) humility and clarity. It reminds us that nature doesn’t always need a new law when old principles can be stretched a bit further. And it refocuses our efforts on leveraging those known principles to improve lives (since if fairness isn’t guaranteed, we’d better actively promote it through social and individual means). Whether or not LoF stands as a new law, the pursuit of this question ensures that we will better understand the hedonic treadmill we all run on, and how far its belt can stretch.

14.5.9 Where we go next:

With adaptation and opponent processes examined at their strongest, we now test a second major rival: predictive coding and free-energy. Chapter 15 asks whether minimizing uncertainty alone already implies the patterns LoF claims.

Chapter 15 — Predictive Coding and Free-Energy

If you follow the last twenty years of brain science, one banner unites many camps: the brain is a prediction machine. In this view, perception is controlled hallucination, action is inference made flesh, and learning is the long, careful work of reducing the surprises the world throws at us. The umbrella framework – predictive coding and the Free-Energy Principle (FEP) – is elegant, mathematically sophisticated, and astonishingly generative. It can help explain why pupils dilate at odd moments, why you feel carsick when expectation and motion disagree, why anxiety tightens when the future feels uncertain, and even why living systems look like they are resisting entropy locally. In this chapter we present the strongest version of that rival program, translate it into plain speech, and then ask the sobering question our project always asks: does this framework, by itself, guarantee fairness, or does it merely tend toward stability?

At its heart, predictive coding says the brain builds hierarchies of expectations (priors) about the world and about the body. Incoming signals are compared against these predictions. Mismatches create prediction errors that travel upward; updated predictions travel back down. Over time, the system learns internal models that make fewer costly errors. The FEP extends this to a unifying principle: organisms act to minimize a quantity called variational free energy, which upper-bounds surprise. You can think of free energy as a running bill for how far off-base your inner story is about what's happening and what will happen next. Action changes the world to fit the model (you turn on a light so the visual input matches your "bright room" prediction); perception changes the model to fit the world (you update your belief to "the room is dimly lit"). Learning slowly reshapes the priors so that future errors shrink.

Affective life fits this picture as well: feelings reflect the precision-weighted balance between what the brain expects and what sensory evidence reveals, especially evidence coming from inside the body (interoception). When predictions are held with high confidence (high precision) and the evidence disagrees, the errors bite hard – you feel alarm, anxiety, pain. When predictions are uncertain and new evidence resolves that ambiguity, you feel relief or even pleasure. Through this lens, many emotional dynamics – habituation, opponent-process rebounds, the soothing power of explanation, the panic of uncertainty – fall out naturally from the math. The active inference extension adds that we do not just wait for the world to calm our errors; we also act, move, and seek information to bring the stream of sensations in line with what our generative model can handle.

So far, this sounds like it could be the whole story. But our project sets a higher bar. The Law of Fairness (LoF) asserts a lifetime constraint on the integral of experienced feeling

– an end-of-stream neutrality that is not merely likely but guaranteed, given certain identity conditions. Predictive coding and FEP, on their own, posit no such global ledger or terminal boundary condition. They describe local moves: given your priors and the current data, you update beliefs and act to reduce expected error. They do not specify that the sum of felt valence must converge to zero by the death of mind. They may tend toward equilibrium in many situations, but they do not promise it.

$$\tilde{L}(t) = \int_0^t HCl(\tau) d\tau$$

$$L(T) = \int_0^T F(t) dt$$

(Here $F(t)$ is our felt affect/valence rate in the life-ledger sense, not variational free energy.) This is why predictive coding and FEP are our strongest rivals and our closest collaborators. They give us mechanistic hooks – precision control, hierarchical message passing, interoceptive priors, control hubs that modulate gain (regions like ACC, rIFG, insula, vmPFC) – which we can actually go measure. They give us falsifiable expectations for how the nervous system will behave when horizons shrink or when the body’s signals are noisy. But left unmodified, they stop at local rationality under uncertainty. They optimize “don’t be surprised,” not “end fair.”

What you’ll get from this Chapter:

- The predictive brain demystified: A clear, plain-language explanation of the predictive coding framework – you’ll grasp how the brain minimizes surprise (variational free energy) and why this theory has gained so much ground in neuroscience.
- Feelings as prediction errors: Insight into how the theory maps happiness and suffering onto prediction errors and precision tuning. You’ll see how emotions can be interpreted as signals about what the brain expected versus what actually happened.
- Active inference vs. fairness: An understanding of where acting to minimize surprise (active inference) naturally creates balance-like effects and where it falls short. We highlight which LoF-like patterns predictive coding can explain (and which it can’t) without adding new assumptions.
- Merging fairness with prediction: A proposed way to integrate the Law of Fairness with a predictive processing model – envisioning LoF as a global constraint layered on the predictive brain (for instance, through horizon-sensitive precision control that ensures neutral closure).

- Decisive experiments: A preview of empirical tests that distinguish a “pure” predictive-coding world from a “predictive coding + fairness” world. You’ll learn what evidence could show that adding a fairness constraint yields better predictions or unique signatures that standard models can’t match.

Subsections in this Chapter:

- **15.1 The Big Idea (Uncertainty Minimization)** - Restates predictive coding’s core logic in accessible terms (with minimal equations) so that the reader sees exactly what the “prediction machine” brain is doing.
- **15.2 Affect as Prediction Error** - Maps feelings onto prediction errors and precision control, showing how pleasure and pain emerge as the felt “shadow” of the brain’s probabilistic bookkeeping.
- **15.3 Active Inference and Fairness Limits** - Describes active inference – how action (not just perception) reduces expected error – and examines where this standard framework can explain LoF-like phenomena and where it cannot.
- **15.4 LoF as a Global Constraint over PC** - Proposes a reconciliation path by layering the Law of Fairness on top of a predictive coding organism. We illustrate how horizon-sensitive precision policies and menu pruning could enforce neutral closure within a predictive brain.
- **15.5 Critical Tests (PC vs. LoF)** - Outlines experiments to distinguish a pure predictive-coding model from a predictive-coding-plus-fairness model. We discuss horizon-contingent narrowing of options, extra variance compression near life’s end beyond medical care, and QS-like residual signals after accounting for other factors.

Where we go next:

We begin with the nuts and bolts. Section 15.1 distills predictive coding’s uncertainty-minimization loop in plain terms, setting up Section 15.2’s careful mapping from precision-weighted prediction error to felt affect.

15.1 The Big Idea (Uncertainty Minimization)

If you strip modern brain theories down to their core, one message remains: living things survive by being less wrong about what happens next. The predictive-coding and free-energy framework makes that message precise. Brains build internal models that generate expectations about incoming signals from the world and the body. Sensory input is compared against these expectations; the mismatch is a prediction error. Big mismatches feel bad or alarming, small mismatches feel safe or satisfying, and learning tries to make the next mismatch smaller.

15.1.1 The Loop in One Paragraph

At each moment, your brain sends top-down predictions to sensory areas and receives bottom-up errors in return. If the light you actually see is dimmer than you predicted, the visual areas send an error signal upward. Two fixes are available: perception can update the internal model to better fit the light, or action can change the world to better fit the model (for example, by turning on a lamp to make the room as bright as you expected). In parallel, the system sets precision – how much to trust a given stream – by adjusting neural gain. High precision makes a given error matter more; low precision lets it slide. *This ongoing dance — predict, compare, correct, and tune precision — keeps an organism near states it can handle.*

15.1.2 A Minimal Equation You Can Read

Predictive coding can be written many ways; we only need the gist.

Prediction error: $\varepsilon = \text{input} - \text{prediction}$.

Precision-weighted error: $\tilde{\varepsilon} = \Pi \cdot \varepsilon$, where Π (pi) is the precision (roughly, inverse variance; the “confidence” assigned to that channel’s input).

Update rule (cartoon): $\text{new belief} \leftarrow \text{old belief} + \text{learning_rate} \times \tilde{\varepsilon}$.

Variational free energy \mathcal{F} is an upper bound on surprise. To avoid symbol collision with our use of $F(t)$ for affect, we’ll write free energy as \mathcal{F} in this chapter. Minimizing \mathcal{F} means choosing beliefs and actions that reduce expected precision-weighted errors over time.

In plain language: *keep the world and the story you tell about it from drifting too far apart, especially where it matters most.*

15.1.3 Why This is Biologically Sensible

There are several reasons such an error-minimizing strategy would evolve in brains. Metabolism: Big errors force costly neural work and frequent corrective actions; energy-hungry brains prefer a manageable, steady error load rather than wild spikes. Learning:

Shrinking recurring errors means future inputs become more predictable; the system harvests regularities and exploits them to its advantage. Behavior: Actions are not only for seeking reward – they also serve to sample the world in ways that keep uncertainty low. We ask clarifying questions, check a door twice, or scan a crowd for a friend’s face, all in service of reducing surprise.

15.1.4 Precision as the Quiet Superpower

Not all errors deserve equal weight. A rustle in the bushes at night gets high precision; the same rustle at noon gets much less. The brain tunes precision by modulating neural gain (think: systems often implicated include the locus coeruleus for overall arousal, the anterior cingulate cortex and right inferior frontal gyrus for cognitive control, the insula for internal bodily signals). Too little precision and you’ll miss important threats; too much precision and you get pathological anxiety, pain amplification, or compulsive double-checking. (Indeed, both therapy and anesthesia can be framed as re-tuning precision: therapy teaches you to down-weight certain prediction errors, and anesthesia globally dampens precision on incoming pain signals.)

15.1.5 How Affect Enters the Picture

Feelings are not extra add-ons; they are summary signals of how well prediction and control are going. A good fit with confident predictions yields feelings of ease, fluency, even pleasure. Being wildly wrong (especially on things deemed important) yields alarm, ache, dread. Relief is the *felt drop* in precision-weighted error when uncertainty resolves. In short, affect can be modeled as a signed, precision-weighted aggregate of prediction errors across channels at a given moment; whether and how such signals can be summed across time into a lifetime ledger is an additional assumption not entailed by predictive coding itself.

15.1.6 From Perception to Active Inference

Prediction errors can be reduced either by changing your beliefs or by changing the world. Active inference formalizes the second path: *we choose actions that we expect will lower future errors*. For example, you pull a blanket over yourself when your interoceptive model predicts “I’m going to be cold”; you ask for clarification when your language model (in your head) predicts you might not understand a conversation. Over time, goals and habits emerge as efficient ways of steering into low-error regions of state space.

15.1.7 A Quick Tour of the Hierarchy

In predictive processing, the brain is often described as a hierarchy of levels. Higher levels encode slow, abstract causes: identities, contexts, long-range expectations,

values and goals. Lower levels encode fast, concrete details: edges and colors in vision, syllables in speech, heartbeats and lung stretches in interoception. Information flow: Errors flow upward while predictions flow downward. Precision is tuned at each level, often discussed in relation to neuromodulators (for example, dopamine is often modeled as signaling volatility and value uncertainty, serotonin as shaping patience and punishment sensitivity, norepinephrine as global arousal, acetylcholine as sensory gain). Each level thus adjusts how “loudly” it shouts about mismatches and how much it trusts the level below.

15.1.8 Where we go next:

With the loop and its minimal equation in place, we now ask how these quantities feel from the inside. Section 15.2 translates precision-weighted error into everyday hedonic terms—why mismatch can sting, completion can soothe, and information can act like analgesia.

15.2 Affect as Prediction Error

If predictive brains survive by being less wrong, affect is how that wrongness feels. In one influential predictive-processing account, emotions can be framed as interoceptive inference—predictions (and their errors) about bodily states (Seth, 2013). In that spirit, we'll treat feelings as tracking the precision-weighted prediction errors described above. In the predictive-coding and Free Energy Principle (FEP) picture, feelings are not decorations on thought. They are the felt summary of precision-weighted prediction error flowing through your interoceptive and exteroceptive channels—essentially, a measure of how far the world (and body) is diverging from what your brain expected, adjusted for how confident you were in those expectations.

The one-line formula, unpacked: Let “input” be what arrives from the senses or body, and “prediction” be what your generative model expected to happen. The basic mismatch is:

Prediction error:

$$\varepsilon = \text{input} - \text{prediction}.$$

Not all mismatches matter equally. The brain encodes a precision Π (often modeled as inverse variance or neural gain) and weights the error by it:

Precision-weighted error:

$$\tilde{\varepsilon} = \Pi \cdot \varepsilon.$$

A tractable modeling assumption is that momentary affect $F(t)$ correlates with a signed, precision-weighted aggregate of prediction errors across channels and levels:

$$F(t) \approx \sum_{c \in \text{channels}} w_c \tilde{\varepsilon}_c(t),$$

with channel weights w_c learned over development. This expresses a correlation or proxy relationship, not an identity claim. For survival reasons, errors in pain or hunger channels tend to receive higher effective weights than minor perceptual mismatches. A positive $F(t)$ corresponds to relatively fluent model-world fit, and a negative $F(t)$ corresponds to salient, high-precision mismatches.

15.2.1 Why the sign flips

Pleasure: Predictions fit inputs with high confidence, or uncertainties are resolved in your favor—your model “catches” the world successfully. Think of the satisfaction of a solved puzzle, a comforting hug that you anticipated, or the relief of pain disappearing right when the medicine was expected to kick in.

Distress: High-precision errors accumulate or persist — your model is loudly wrong in ways that matter. Sudden loss, social rejection, an escalating physical threat: these all violate strong predictions and generate a torrent of error signals (hence intense negative feelings).

Relief: Not simply having a low error, but a drop in error. Relief is the palpable change when ε falls dramatically after being high. That's why turning off a blaring car alarm yields a wave of pleasure beyond the baseline quiet – the silence is golden only because the noise was so high before.

15.2.2 Interoception first

A large share of our affective life comes from internal bodily prediction errors. The insular cortex (from posterior to anterior) integrates bodily sensory signals; the anterior cingulate cortex (ACC) tracks control effort and conflict; the ventromedial prefrontal cortex (vmPFC) and orbitofrontal cortex integrate value under uncertainty; the locus coeruleus (noradrenergic center) modulates global precision; serotonergic systems adjust our tolerance for waiting and our sensitivity to loss. In this architecture:

Interoceptive prediction errors (hunger, pain, breathlessness, fatigue) are heavily precision-weighted and thus disproportionately influential on how you feel. The body's signals demand attention when they deviate from prediction.

Exteroceptive errors (sights, sounds, external events) only sting strongly when they imply bodily risk or social threat. The precision on these is often gated by context – a sudden rustle is weighted more at midnight in an empty alley than at noon in a park, and a stranger's glance carries more weight if you're alone than if you're with friends.

15.2.3 Everyday corollaries that match experience

Habituation and hedonic adaptation: Repeated, predictable events shrink either ε (the error) and/or Π (the precision given to that input), so feelings gravitate back toward neutral unless novelty or stakes ramp up again. This lens is often presented as consistent with classic discussions of hedonic adaptation. Brickman (1978) is frequently cited in this context, suggesting partial adaptation in reported happiness following major positive and negative life changes. Such results align with the predictive-coding view: as surprising events become familiar, their error (and emotional impact) diminishes. This aligns with the familiar tendency to get “used to” things.

Why information soothes: An explanation reduces uncertainty about what's happening, which in effect reduces precision Π on potential errors. Even before the world itself changes, just knowing why something happened can lower the error load you feel.

Why ambiguity hurts: If you assign high precision to all the bad possibilities you can't rule out, not knowing (ambiguity) keeps ε elevated. Sustained not knowing — waiting for results, living with uncertainty — often translates into sustained negative affect purely because of those high-precision, unresolved prediction errors.

Why completion feels good: Finishing a task or closing a chapter in life removes a whole set of “queued” error sources. This produces a drop in cumulative ε , which we experience as satisfaction or relief. Completing things literally lightens the load of prediction errors the brain is juggling.

15.2.4 From momentary affect to a daily curve

If you plotted $F(t)$ over the course of a day, you'd see a jagged curve: morning stress spikes, a friendly lunch dip into positive territory, a tough meeting punching it downward, an evening run pulling it back up, a kind text giving a small positive bump, and so on. Some of these error “pulses” are brief and intense, others are long and dull. In our broader framework, the life ledger is essentially the integral of this curve over an entire lifetime. Predictive coding explains the shape of the curve locally (why each pulse goes up or down). The Law of Fairness asks whether the integral of that curve obeys a global neutrality constraint by the end of life. In other words, PC describes the day-to-day fluctuations; LoF inquires about the net sum of those fluctuations over the long run.

15.2.5 How the brain adjusts the “volume” on feeling

Affect isn't just raw mismatch; it's mismatch under a chosen confidence policy. The brain dynamically controls how “loud” errors are felt:

Arousal systems (norepinephrine): These boost precision globally during moments of threat or challenge, making every error signal hit harder and making corrective learning faster. (This is why a scare can be a potent learning event.)

Contextual gating (prefrontal cortex and hippocampus): Your brain assigns higher precision to errors in high-stakes contexts (e.g. driving on an icy road, where even small deviations matter greatly) and lowers precision in safe contexts (lounging on the couch, where minor surprises are no big deal).

Maladaptive settings: Chronic anxiety or chronic pain often involve persistently high precision on particular channels. The result is that even ordinary inputs generate a lot of felt error — normal sensations or minor worries are experienced as significant distress because the “volume knob” (precision) is stuck too high on those channels.

15.2.6 Clinical resonance

Depression: In predictive-coding terms, this can be framed as a lowered expectation of control (the person's model predicts "nothing I do will help") combined with high precision on negative predictions. This means even small negative mismatches feel heavy, and positive surprises barely register because predicted pleasure has very low precision. The outcome is anhedonia (nothing feels good) and persistent distress.

PTSD: In predictive-coding terms, traumatic experiences can leave priors with unusually high precision. As a result, benign cues produce large ε – for example, a backfire sounds certainly like danger. Therapy can be seen as a process of gradually retuning those precisions and predictions: the world is relearned to be less automatically dangerous, reducing those massive error signals.

Placebo/Nocebo effects: Beliefs (whether in a treatment or an expectation of pain) change the brain's predictions and perceived precision about incoming signals. This can alter experienced pain or relief without changing the actual sensory input at all. The placebo pain relief isn't "fake" – it's the brain literally predicting less pain and thus reporting less error.

15.2.7 Why this is good science

The predictive coding account of affect is mechanistic (it maps to identifiable neural circuits), quantitative (it gives a signed and weighted error value at any moment), and testable (it predicts what should happen if we manipulate precision or surprise experimentally). It links perception, action, learning, and feeling in one elegant loop – with no extra "feeling module" or magical process needed. In this view, emotions emerge naturally from the same computations that drive perception and behavior.

15.2.8 Where it stops, and where the Law of Fairness begins

Crucially, nothing in precision-weighted error minimization forces lifetime fairness. One person could live in conditions where negative, high-precision errors dominate for decades; another could enjoy a life of fluent, high-confidence matches with few errors. Predictive coding explains strong tendencies (homeostatic regulation, adaptation toward baseline) but does not guarantee a neutral integral of suffering and joy over the lifespan. That gap is exactly what the Law of Fairness (LoF) proposes to fill: LoF posits a global constraint that shapes the menu of admissible thoughts and actions as a function of the current ledger and time horizon. In practice, this would show up as horizon-sensitive precision policies and control thresholds that tilt choices toward compensable moves (like repair, rest, reconciliation) when neutrality is at risk.

15.2.9 Joint predictions (PC-alone vs. PC + LoF)

We can now summarize differences in what pure predictive coding would predict versus a predictive-coding-plus-fairness model:

Horizon sensitivity: Predictive coding alone predicts that precision tuning depends only on local stakes and uncertainties. PC+LoF predicts an additional tightening of control as a hard horizon (e.g. end-of-life) approaches – a tightening specifically aimed at boosting reparative, reversible options even when immediate utility or risk is the same. In other words, as time runs out, the system puts a special premium on actions that can “close the ledger.”

Terminal variance compression: Predictive coding can explain some reduced variance in affect near death through sedation, routine, and reduced stimuli. PC+LoF predicts cross-modal compression of hedonic variance as the end draws near – a systematic shrinking of ups and downs – after controlling for the effects of medication and care. Moreover, LoF expects content-specific shifts (e.g. a tilt toward closure-related emotions like forgiveness or acceptance) that pure predictive regulation wouldn’t necessitate.

Dream counterweights: Predictive coding treats dreams as offline model updates or rehearsals, with no requirement for any hedonic pattern. LoF uniquely predicts systematic valence counterbalancing in dreams: when waking life is skewed strongly negative one day, that night’s dreams will tend to be positive or reparative; after an unusually positive day, dreams might turn more sobering or negative. In short, LoF suggests dreams serve as hedonic counterweights to maintain the ledger, which standard predictive models do not require (beyond processing salient events).

15.2.10 Minimal lab tasks to make this concrete

Precision cueing task: Experimenters manipulate the expected reliability of a sensory stream (e.g. sometimes you’re told a signal is very trustworthy, other times not) and measure how your affect tracks these changes. Pure PC predicts your feelings will track precision-weighted prediction errors. LoF makes an extra prediction: if we introduce a “horizon” cue late in the task (say, hint that the experiment is about to end), you’ll show extra pruning of low-compensatory, non-repair choices – an urgency to resolve things – beyond what precision and reward alone would dictate.

Ambiguity relief task: Induce uncertainty about an outcome and then resolve it without actually changing anything material (for example, you’re unsure if you won a reward; later we simply tell you the outcome was neutral). We measure the relief as the drop Δ in ε . LoF predicts stronger relief when the person’s running ledger is in the red (negative) and

the time horizon is short – because clearing uncertainty when you’re already down and time is running out would have outsized emotional benefit.

Counterfactual regret assay: Present choices that have equal immediate payoff but differ in their future “repairability.” For instance, Option A gives a moderate reward but no chance to fix mistakes later; Option B gives the same reward with a built-in opportunity to reverse it if it turns bad. Predictive coding (and standard decision theory) predict you’d feel about equally good choosing either, since immediate utility is matched. LoF predicts that as the horizon shrinks, people will show a preference for the option with future reparability and feel less regret choosing it – essentially valuing an action for its impact on the long-term ledger, not just its instant reward.

15.2.11 How to read your day through this lens

The tightness in your chest before sending an angry message isn’t mystic morality—it’s high-precision prediction error about near-future social fallout. The deep ease after finishing a long-avoided task comes from a drop in your error baseline—not just a to-do check mark, but your brain registering that a lingering discrepancy is gone. When time feels short (a deadline, health scare, birthday), “right” options seem to shine while distractions fade. Predictive coding explains part of that (risk aversion and focus under time pressure). LoF adds a global fairness tilt—your mind nudging you toward what matters for closing the ledger.

15.2.12 Failure modes of the affect-as-error account and how we guard against them

All-purpose vagueness: If any feeling can be explained post hoc as “prediction error,” the theory becomes unfalsifiable. We counter this by preregistering experiments that manipulate precision or surprise in specific ways and by making *a priori* predictions about expected brain and behavioral changes (which regions activate, how reaction times shift).

Confounds with utility: One might argue pleasure is “just reward,” not error resolution. To separate them, we hold immediate reward constant but vary reparability and timing. If equal payoffs yield different affective outcomes depending on whether later repair is possible, that supports error-resolution value beyond raw utility.

Circular measurement: We must avoid explaining self-reported feelings purely in terms of those same self-reports (tautology). Thus, we insist on convergent measures of affective prediction error: physiological signals (pupil dilation, heart-rate variability), neural markers (activity in vmPFC, ACC, insula, etc.), behavioral choices (persistence vs. quitting, “checking” behaviors), and even dream content analysis. If all these align, we

have much stronger evidence that we're measuring something real about prediction errors and not just re-labeling someone saying "I feel bad" as "high error."

15.2.13 Where we go next:

Where this leaves the reader: You can carry one clean idea forward: Feeling = how wrong you are, weighted by how much you care about being right (in that moment). That's a powerful, disciplined lens on human experience. The Law of Fairness does not ask you to throw this lens away; it simply asks whether the sum total of those weighted wrongnesses and rightnesses is subject to a deeper bookkeeping rule across an entire conscious life. In the next section (15.3), we explain why a brain that excels at minimizing uncertainty can still leave some lives fundamentally lopsided — and why viability is not the same as fairness.

15.3 Why Viability ≠ Fairness

A living system can be exquisitely good at staying alive while being catastrophically bad at being fair. Viability is a biological success criterion: maintain homeostasis, avoid hazards, seek resources, reproduce if possible. Fairness, in our usage, is a moral-experiential criterion: does the integral of felt experience across a conscious life settle to neutral (neither a surplus of pleasure nor a deficit of pain) by the death of mind? These two notions can correlate, but they are not the same variable and they need not move together.

15.3.1 Robust but lopsided lives

Many organisms adapt to relentless adversity. Their brains become extremely efficient at prediction and control under scarcity or stress. They survive (and even meet ethological definitions of “thriving”), yet they accumulate decades of negative affect. Viability succeeds; the hedonic ledger does not self-correct in any obvious way.

15.3.2 Pleasant short lives

Other individuals might experience prolonged safety, comfort, and reward but die young. Their viability fails (they don’t survive long), while their affect to date might be net positive. Again, the survival metric and the hedonic accounting diverge.

15.3.3 Adaptive indifference

Predictive models can reduce the felt weight of error by lowering precision rather than by improving the world. In plain terms, numbness is cheap. An organism can blunt its sensitivity (through learned helplessness or habituation) and thereby feel less wrong without actually correcting anything. A dulled meter is not the same as a paid debt — a life’s burdens can remain unclosed even if the person reports “I’m fine” because their system has simply stopped registering the imbalance sharply.

15.3.4 Local optima, global harms

Behaviors that minimize immediate variational free energy \mathcal{F} (uncertainty/surprise) can entrench long-run suffering. For example, avoidance coping will reduce near-term prediction errors (“I feel safe now because I didn’t challenge myself or face the issue”), but it can foreclose on opportunities for repair or growth, leading to worse outcomes later. The organism remains viable — it’s finding short-term optima — while fairness (balance of life outcomes) drifts further off.

15.3.5 Why predictive success does not guarantee a neutral ledger

Predictive-coding and free-energy accounts explain how brains keep error manageable. They optimize moment-to-moment fit under current constraints. But they do not specify any global path constraint over an entire lifetime's integral of affect. Two gaps are critical:

Temporal scope gap: Homeostatic regulation, reinforcement learning, and uncertainty minimization operate on timescales from milliseconds to months. The Law of Fairness claims a boundary condition at the scale of a whole life. Nothing in standard “viability” principles forces a neutral terminal balance — they weren't designed with that scope in mind.

Objective gap: Minimizing surprise keeps organisms operationally viable even if their world is predictably unpleasant. A factory can run smoothly and efficiently while churning out miserable products. Likewise, a brain can achieve a steady state in which it expects and accepts suffering as the norm — it's minimizing surprise (because pain is no surprise), but that life is not fair in the hedonic sense. In short, a system can perfectly well optimize “don't be caught off guard” while generating sorrow just as efficiently as joy.

15.3.6 Analogy – temperature vs. entropy

Viability is like keeping your body at 37 °C – maintaining an operational equilibrium. Fairness is like constraining the total entropy change over an entire thermodynamic trajectory. You can hold temperature constant while dumping entropy elsewhere; analogously, an organism can maintain homeostasis while exporting the “cost” to one particular stream of experience (e.g. constant emotional pain). One constraint does not entail the other.

15.3.7 The practical upshot

Therapies that improve viability may leave fairness untouched. An anxiolytic drug that lowers precision will make someone feel less anxious (dampen the error signals) without necessarily closing any of the underlying “debts” in their life. High-dose analgesia can eliminate pain in the moment but might not repair the social or psychological wounds that contributed to that pain.

Environments that boost viability can mask unfair ledgers. Providing stable employment, food security, and housing will reduce the spikes of error an organism experiences (a very good thing for survival and basic well-being). But by themselves, these don't ensure that a person who has suffered deeply will also encounter the compensatory positive experiences needed for a neutral lifetime ledger. Survival conditions can be excellent while the hedonic ledger remains deeply negative (or positive).

15.3.8 Both metrics matter

Ethical and scientific evaluations need to track both. A clinical trial might show great outcomes on symptom reduction or survival time, yet those metrics could obscure whether the person's life-integrated affect is balanced or not. We can imagine an intervention that keeps patients calm (low precision, low response) but effectively leaves them with an "unpaid" emotional debt — something that would never show up if we only measure short-term stress or vital signs.

15.3.9 Concrete contrasts in the lab and clinic

Horizon tasks: In experiments where participants face a short time horizon (e.g., a scenario framed as their "last chance" vs. one of many future opportunities), LoF predicts a distinct tilt toward reparative, reversible choices — even when immediate payoffs and prediction errors are matched. In contrast, standard viability accounts (predictive coding, RL, etc.) might predict increased caution or urgency under short horizons (due to risk aversion or higher stakes), but not a special premium on closure or reparability per se. If we see people disproportionately choosing the option that "keeps the ledger clean" when time is short, that's something beyond generic risk-sensitivity.

Terminal contexts: In end-of-life settings under good palliative care, predictive accounts expect overall affective variability to drop (the person is sedated, routines are controlled, external surprises are minimized) and for absolute affect levels to hover near neutral simply because strong emotions are dampened. LoF adds a more specific signature: cross-modal variance compression coupled with closure content. That means not only is variability lower, but whatever feelings do emerge skew toward meaningful closure-related themes (forgiveness, reconciliation, relief), beyond what sedation and routine care would produce. Essentially, if people near death consistently experience a narrowing band of affect centered on peaceful neutrality and resolving experiences, LoF is in play; if they instead exhibit random or widening variation (even if small) once medicine is accounted for, then LoF is not needed to explain it.

Dreams: Predictive coding treats dreams as the brain's offline model-fitting and memory consolidation time. There's no requirement that dream content systematically compensate for waking mood; dreams should reflect recent salient stimuli or unresolved conflicts, etc., but not necessarily invert their valence. LoF, however, predicts that dreams will function as counterweights: after days that are heavily negative, dreams will skew positive or restorative; after overly positive days, dreams may introduce negative or sobering elements. In other words, LoF expects an allocation policy in dreams aimed at

evening out the ledger, whereas standard theory sees dreams more as random “replay” or problem-solving without that specific balancing agenda.

15.3.10 Edge Cases: Learned Helplessness vs. Neutral Closure

In learned helplessness (a classic viability adaptation to escape futility), an organism lowers precision and suppresses action. This can stabilize physiology (conserve energy, avoid futile attempts) – a win for viability – while effectively etching a large negative experience into memory and identity (“life is pain and there’s nothing I can do”). The person survives but with a deeply negative ledger. LoF forbids ending a life in that state of unresolved deficit; it predicts that if an individual reaches late life with such an imbalance, circumstances will tend to change in ways that reopen possibilities or reweight experiences (perhaps through unexpected social support, personal insight, etc.) to prevent a permanently lopsided outcome. If we observe many people dying in a state of profound learned helplessness without any final-hour shift, that undercuts LoF.

Ascetic equanimity: Some contemplative traditions cultivate a state of equanimity – extremely low prediction error across all contexts. The person feels unperturbed by things that would normally cause stress. If this equanimity comes from genuinely resolving attachments and closing emotional “accounts” (i.e. they’ve achieved a kind of comprehensive repair of desires and fears), LoF is satisfied. But if it comes from globally dialing down precision or detaching from outcomes without actually addressing any debts (a kind of induced apathy), then the neutrality is superficial or “pseudo-neutral.” The distinction is empirical: we’d look at whether the ascetic individual still has unreconciled relationships or unaddressed traumas versus truly having closed those loops. LoF would predict that real neutrality correlates with evidence of reconciliations and completions, whereas simple indifference without such acts might not hold at life’s end.

Palliative terminal lucidity: Hospice workers often report cases where a dying patient suddenly becomes clear-minded, energetic, and engages in meaningful goodbyes or reconciliations shortly before death (sometimes called the “last rally” or terminal lucidity). From a standard medical view, these can seem like anomalies or random brain events. Under LoF, they are diagnostic: when the horizon shrinks to near-zero, the admissible menu of brain states widens temporarily for crucial closure acts, even if by purely viability criteria the brain should only be deteriorating. In other words, LoF would frame terminal lucidity as the control system giving one last push to balance the ledger (allowing final conversations, forgiveness, expressions of love) despite the overall collapse of viability.

15.3.11 Why neutrality is stronger than “tendency toward baseline”

It's important to distinguish LoF from the well-known idea of hedonic adaptation (“most people tend to drift back to a baseline mood after big life events”). Hedonic adaptation is a statistical tendency, not a guarantee – it speaks to group means and typical outcomes, but says nothing about the tails. A population can show a reliable mean reversion while some individuals end up in chronic despair or lifelong euphoria. A law must bind even the tails. LoF claims the binding happens at the terminus of a conscious life, not at daily or yearly scales. It tolerates long deviations – even a lifetime of extreme imbalance – and insists only on eventual neutral closure (given the organism remains its same identity until the end).

15.3.12 A minimal mathematical picture

Let $F(t)$ be precision-weighted affect at time t , and let $L(T) = \int_0^T F(t) dt$ be the life ledger accumulated up to the horizon T (the end of life). Viability aims to keep $F(t)$ bounded and typically near 0 through continuous adaptation; this ensures that the integral $L(T)$ exists (doesn't diverge to infinity) but places no requirement on its value. LoF asserts a terminal condition: $\lim_{\{t \rightarrow T^-\}} L(t) = 0$ (within tolerance K), enforced by shaping of the admissible set $A(t; \bar{L}, H, C)$ – essentially, as the shadow price λ_t rises with a shrinking horizon H , the system becomes increasingly conservative about letting the ledger drift. In practical terms, “within tolerance K ” might mean that in a person's final stretch of life, their standardized affect is centered within ± 0.15 (nearly neutral), with a slope no steeper than ± 0.05 per day and variance compressed to $\leq 80\%$ of earlier levels. Viability, on its own, can be perfectly satisfied with $L(T) \ll 0$ or $L(T) \gg 0$ (a life ending in deep deficit or surplus). LoF forbids both extremes.

15.3.13 Policy implications if you accept the distinction

Metrics: Track two dashboards for well-being – one for survival/functional outcomes and one for ledger movement (longitudinal measures of aggregate affect, e.g. using our Hedonic Composite Index over time). Do not assume that suppressing symptoms or keeping someone stable is equivalent to “paying off” their experiential debts.

Design: When horizons tighten (terminal illness, old age, even shorter crises), prioritize interventions that maximize reparability and reversibility. In healthcare or counseling, this means create opportunities for reconciliation, forgiveness, making amends, expressing final truths – these are high-“Φ” (feasibility-of-compensation) acts. Likewise, keep as many channels open as possible (social, creative, sensory channels) for positive experiences to occur.

Evaluation: Judge societal and institutional success not only by how many bodies are kept alive and functional, but by how many lives approach neutral closure. A hospital could be excellent at keeping people alive (viability) while failing to ensure those lives end in resolved, balanced states. Both aspects should be measured.

15.3.14 What would falsify the fairness claim (while leaving viability intact)

A law earns its keep by stating what would count against it. Here are clear patterns that, if observed under good conditions, would refute LoF even though predictive brain theory stays unchallenged:

Horizon-invariant menus: If, after rigorous controls, people's options and urges near the end of life show no tilt toward closure moves (e.g., no spike in desire to reconcile, no shift toward meaningful or reparative activities) – essentially if the admissible set of late-life actions $A(t)$ does not narrow toward high- Φ options – then LoF loses a key behavioral signature.

Terminal dispersion: If well-measured end-of-life affect shows expanding (or at least not compressing) variance, even after controlling for things like sedation levels, pain relief, counseling, etc., then the LoF prediction of final hedonic narrowing is violated. In simple terms, if as people approach death their emotional states become more erratic or remain all over the place (instead of converging to neutrality), LoF is in trouble.

Null counterweights: If a person experiences a persistently skewed emotional life (say, months of continuous sorrow or mania) and yet we do not observe any systematic compensatory dreams, daydreams, or micro-experiences injecting the opposite valence, then LoF's story about the brain allocating counterweights is weakened. For example, if an individual with prolonged, well-documented negative affect (with adequate measurement density and open compensatory channels) shows no statistically detectable increase in opposite-valence micro-experiences relative to baseline or matched controls, that would count against the necessity of a fairness mechanism.

15.3.15 How to read your own life through this lens

One practical takeaway is to periodically ask two different questions. First: "Am I viable?" That is, am I essentially surviving and functioning – eating, sleeping, learning, connected to others? Second: "Am I closing the ledger?" In other words, what emotional or relational debts might I be carrying (unsaid apologies, unresolved conflicts, unfulfilled aspirations, untreated pains), and what steps are available to start moving those toward zero? The first question keeps you here (alive and stable). The second decides what *here* adds up

to in the end. LoF encourages not just a healthy life, but one that tends toward completeness.

15.3.16 Where we go next:

We've now argued conceptually that viability and fairness can dissociate; next we need empirical ways to tell them apart. The following section (15.4) sketches research tests that separate predictions all frameworks agree on (overlap) from those where LoF and standard predictive/homeostatic theories would diverge (independence tests). These are crucial, because they let data arbitrate between "this is just the brain reducing surprise" vs. "this is the brain obeying a fairness constraint."

15.4 Research Notes: Overlap vs. Independence Tests

This section separates predictions LoF shares with mainstream theories (overlap) from those it uniquely makes (independence). Overlap tests show that LoF sits comfortably within the successful core of existing science (it doesn't contradict known phenomena); independence tests are designed to decisively pit LoF against alternative frameworks. For each study idea below, we outline the design, key variables, predicted pattern, and clear fail conditions.

These tests by themselves won't prove LoF, but they ensure that LoF is consistent with well-established findings. LoF should at minimum replicate the known successes of adaptation, opponent processes, and predictive regulation.

15.4.1 Adaptation after shocks

Design: Longitudinal pre-post tracking of individuals after major life events (job loss, breakup, illness, etc.), with daily HCl (hedonic composite) ratings, autonomic markers, sleep quality, activity levels, and social contact logged for ~120 days surrounding the event.

Prediction (overlap): A partial return toward each person's baseline level of well-being within weeks to months after the event, with the speed and extent of recovery depending on context (e.g. faster rebound with strong social support or if the event was expected).

Why it matters: This shows that LoF allows hedonic adaptation dynamics – the fact that people tend to drift back toward baseline after shocks is not at odds with eventual neutrality. LoF doesn't claim you feel zero every day, only that the integral might zero out at the end; interim rebounds are expected. Demonstrating adaptation is thus a compatibility check.

Fail condition: LoF would be in trouble if it somehow required a monotonic daily drift to zero (which it doesn't). This test primarily verifies that LoF can coexist with empirical hedonic adaptation: if people never adapted at all, LoF would be hard-pressed to describe reality, but decades of data show they *do* adapt.

15.4.2 Opponent-process rebounds

Design: Laboratory inductions of acute positive and negative affect (for example, a cold pressor pain vs. a sweet taste reward, or a jump scare vs. a soothing massage). Measure mood and physiological responses for ~60 minutes post-stimulus.

Prediction (overlap): Clear opponent-process rebounds toward baseline. A painful stimulus is followed by a rebound of relief or calm; an intense pleasure is often followed

by a mild “crash” or return toward neutrality. The magnitude of the rebound scales with the intensity of the initial stimulus (stronger stimulus \Rightarrow stronger opposite after-effect).

Fail condition: None specific to LoF – this is a basic phenomenon it happily includes. If opponent-process effects didn’t exist, it would actually undercut much of the empirical basis that inspired LoF. So here we’re mainly checking that LoF’s framework can smoothly accommodate these well-known after-effects (it can, since they can be seen as short-horizon fairness microdynamics).

15.4.3 Predictive homeostasis under uncertainty

Design: Decision-making tasks where environmental volatility is manipulated. For example, a learning task with periods of stable vs. wildly changing reward probabilities. Measure surprise responses (e.g. prediction error signals, pupil dilation as an arousal proxy), learning rate adjustments, and perhaps neural responses in areas like ACC (which tracks volatility).

Prediction (overlap): Participants show appropriate precision-weighting adjustments: during high-volatility periods, we expect larger surprise responses and dynamic adjustments of learning rates (since predictions are less reliable); in stable periods, smaller surprise responses and more gradual learning. Pupilometry should track unexpected uncertainty (dilating with surprise spikes), and ACC should register volatility signals. These are straight predictions from predictive coding/FEP that LoF would also predict, since LoF doesn’t throw out basic brain function.

Fail condition: LoF does not hinge on any one of these immediate effects (they are more foundational). If these failed, it would shake the entire predictive coding framework, not just LoF. In short, a failure here undercuts the common ground on which LoF is built, rather than LoF itself specifically.

15.4.4 Independence tests (where LoF must diverge)

Each of the following studies contrasts LoF’s predictions with those of pure predictive coding/FEP or standard RL-homeostasis accounts that lack a lifetime-neutrality constraint. These tests are designed to reveal patterns unique to LoF. Success on these would strongly support LoF; failure (especially if the rival theories predict outcomes better) would favor the rivals.

15.4.5 Horizon-tilt in admissible menus (behavioral)

Design: A within-subject experiment that manipulates the participant’s subjective time horizon H. Use robust framings to induce a “short horizon” mindset vs. a “long horizon” mindset – for example, telling participants a decision task is their last chance at

something versus part of a longer series, or using mortality-salience primes (prompts that make people think about an imminent ending). In each condition, present choice sets where options are carefully yoked on immediate payoff and risk, but differ in their reparability/closure value. (For example, two investment choices with equal expected return and risk, but one is reversible/cancellable and the other is irrevocable.)

Measures: The proportion of choices that favor high- Φ (high feasibility-of-compensation) options vs. low- Φ indulgent or one-way options; decision response times and drift-diffusion model parameters (to gauge internal conflict/inhibition); self-reported urge or comfort with each choice; and any post-task regret or relief.

LoF prediction: As the horizon H shrinks (short-horizon condition), people's admissible set $A(t)$ effectively narrows and tilts toward reparative, easily reversible choices. We expect a Horizon \times Reparability interaction: when thinking in "last chance" terms, participants will choose the closure-friendly option significantly more often than when they feel they have plenty of future. Decision-making might also slow down for low- Φ (hard-to-repair) options under short horizon (reflecting extra internal veto pressure), and we might see more frequent changes-of-mind or hesitation on those options.

Rival prediction: The main thing standard theories might predict under a short horizon is increased risk aversion or urgency in general. But they would not predict a specific preference for reparable vs. non-reparable as such. Any shifts in choice should be explainable by generic factors like impatience (discounting the future heavily) or loss aversion – not a targeted tilt toward closure. In other words, if an agent is rational and short-sighted, they might avoid high-risk or delayed payoff options, but they shouldn't systematically care about "Can I undo this later?" unless some learned utility is attached to it.

Fail condition for LoF: Across well-powered replications, if we find no Horizon \times Reparability interaction – i.e., people's choice proportions for high- Φ vs. low- Φ options don't change with short- vs. long-horizon framing (after controlling for any trivial time preference effects) – then LoF's unique behavioral prediction is falsified. It would mean a short time perspective doesn't inherently push people toward fairness-preserving options beyond what standard decision factors predict.

15.4.6 Neural shadow price λ_t (systems-level)

Design: An fMRI experiment (potentially paired with pupillometry for arousal/precision) conducted during the decision task from 15.4.5. Use pre-registered regions of interest (ROIs) based on our model: e.g. vmPFC/OFC for value representation, ACC and rIFG for cognitive control (and possibly inhibitory "braking"), and anterior insula for interoceptive

surprise or “gut feeling.” Build a computational model to analyze the fMRI data that includes standard decision variables (utility, risk, conflict, recent prediction error, etc.) plus a Φ term that indexes each option’s expected contribution to eventual ledger neutrality (some measure of its “closure” value). We also include a formal shadow-price variable $\lambda_t \propto H^{-1}$ (basically an urgency signal that grows as remaining horizon H shortens).

LoF prediction: We expect to find a QS-residual in the neural data – meaning, after accounting for utility, risk, and uncertainty, there remains systematic neural activity corresponding to the Φ term (the fairness or compensatory value). Specifically: vmPFC value signals will carry a component encoding the option’s Φ (not just its reward); ACC and/or rIFG will raise decision thresholds or response caution specifically for low- Φ (hard-to-compensate) options as the horizon shrinks; and the insula may reflect an “admissibility” weight (sometimes denoted ω) that flags options which threaten ledger closure. All these effects are hypothesized to be modulated by the horizon – operationalized as a parametric λ_t signal (increasing as H decreases) interacting with Φ -related regressors in pre-registered ROIs. We might see, for example, ACC activity correlating with λ_t when suppressing an impulsive low- Φ choice. The key is that there’s extra, structured brain variance explained by “is this choice keeping the ledger on track?” beyond standard factors.

Rival prediction: In a purely predictive coding or reinforcement framework, value signals in vmPFC should track known quantities like immediate reward, probability, uncertainty, etc., and any additional variance is just noise or idiosyncrasy. ACC/rIFG should reflect conflict or difficulty as typically modeled (perhaps more conflict when an option has lower expected value or higher risk, but nothing to do with Φ per se). In short, there would be no need for a horizon-scaled fairness component: once you’ve included utility, risk, and perhaps a generic time pressure, the brain patterns should be fully explained. Any consistent residual pattern (like “always lights up for irreversible choices at end-of-game”) should not appear unless LoF is true.

Fail condition: If, after very careful modeling and nuisance regression, we find no neural residual corresponding to Φ – i.e., all observed brain activity in those ROIs reduces to classical variables (utility, risk, conflict, etc.) with nothing left that correlates with our ledger/fairness metrics – then LoF has no distinctive neural support. If vmPFC has no extra “closure value” signal, and ACC/rIFG show no special Horizon \times Φ interaction, it means the brain isn’t doing anything beyond standard predictive decision-making. This would significantly weaken the case for LoF as a separate principle.

15.4.7 Dreams as counterweights (sleep lab)

Design: A controlled sleep laboratory study over (say) 14 consecutive nights. Each participant experiences alternating “skew” conditions during the day: on odd days, we impose a negative skew (high stress, high error load day) via difficult tasks, induced conflicts, or other stressors; on even days, we create a positive skew (buffered, low stress day) with enjoyable activities, praise, and support. Importantly, overall sleep conditions are held constant. Each night, we perform REM sleep awakenings (waking the person after REM periods) and collect dream reports immediately (to be recorded and later coded). Dream content is blind-coded by independent judges using a predefined dictionary for emotional tone and themes (looking for positivity/negativity, themes of restoration, loss, social bonding, etc.). We also measure morning mood (HCI) and next-day affect to see if dreams influenced carry-over.

LoF prediction: Valence inversion in dream content relative to the prior day’s skew. After a heavy negative day, that night’s dreams should, on average, be more positive, comforting, or constructive (e.g. containing themes of solving problems, receiving help, retrying something successfully, or otherwise countering the day’s stress). After an overly positive/easy day, dreams might introduce challenges or more sobering elements (not nightmares per se, but perhaps anxiety themes or reminders of unresolved issues). The magnitude of this inversion should correlate with how unbalanced the day was – a very bad day sparks very uplifting dreams, and vice versa – and critically, the dreams should have an adaptive effect: participants with pronounced counterweight dreams should show a reduction in next-day emotional asymmetry (e.g., if Day 1 was awful but the Night 1 dream was strongly positive, the person’s Day 2 mood should rebound more). LoF frames this as the brain allocating relief/repair during off-line hours to prevent sustained imbalance.

Rival prediction: Standard theories (J. Allan Hobson’s activation-synthesis or even modern predictive processing accounts of dreaming) suggest dreams are for processing salient experiences, integrating memory, or exploring possible futures – but they do not require an inverse relation to the previous day’s valence. If anything, one might expect continuity (stressful days yield stress dreams, happy days yield pleasant dreams) or just a mix heavily influenced by salient events regardless of valence. If we control for heightened arousal or salience (e.g. both very good and very bad days are “emotionally charged”), a predictive coding view could say both will lead to more vivid dreams, but not specifically opposite-valence content. No systematic ledger-dependent inversion is expected once factors like general arousal and REM fragmentation are accounted for.

Fail condition: After meticulous blinding and coding, if we find no reliable valence inversion – e.g., dreams simply mirror waking emotions or vary idiosyncratically without the hypothesized pattern – then LoF’s idea of dreams as a deliberate compensatory mechanism is unsupported. If dreams consistently just reflect the most salient concerns (traumatic day → nightmares about trauma, happy day → nice dreams) and not the opposite, then the “counterweight” interpretation fails. A particularly damaging outcome for LoF would be if even highly skewed days do not produce any balancing effect in dreams across many participants and nights.

15.4.8 Terminal variance compression (hospice telemetry)

Design: An observational cohort study in hospice or end-of-life care settings. Participants (with consent, of course) are patients in their final weeks of life. We collect high-frequency data for, say, their last 60 days: regular subjective affect ratings (if possible) or proxy ratings via experience sampling (HCI from brief mood surveys), caregiver logs of mood and significant events, passive sensing of activity/rest patterns, etc. We also note any “reconciliatory” opportunities (visits from family, conflict resolutions, spiritual counsel) and medication logs (especially sedatives, pain meds). The goal is to see what happens to the variance and content of affect as the time horizon H_t approaches 0, under real-world conditions.

LoF prediction: As the horizon contracts (i.e., in the last weeks and days), two things happen: (1) Affective variance compresses – the person’s emotional state (even if still oscillating) stays within a narrowing band closer to neutrality. Extreme highs and lows become less frequent (beyond what would be expected just from medication or inactivity). And (2) content tilts toward closure – a disproportionate number of emotional or interpersonal events revolve around forgiveness, farewell, making amends, meaning-making, and relief from burdens. Even micro-events like a lucid moment might be filled with saying “thank you” or “I’m sorry” rather than random topics. LoF specifically predicts that this pattern appears above and beyond what standard care practices induce. For example, good palliative care obviously reduces pain (and thus emotional swings from pain), and sedation reduces reactivity; we would control for those factors. LoF expects an extra compression of variability and a thematic skew toward closure that cannot be fully explained by “they gave the patient morphine and so they felt calm.” Also, LoF predicts the occasional “burst” of meaningful activity – short windows of lucidity or energy (often noted anecdotally) – which are enriched in closure acts (final conversations, rallying to settle something) more than would be expected by chance.

Rival prediction: A standard homeostatic or predictive account would expect that as death approaches, yes, variability might drop – but mainly due to diminished

responsiveness (the person is tired, medicated, disengaging) and well-managed comfort. In other words, any neutrality at the end is a byproduct of comfort-care and the body's shutdown, not an active push for hedonic balance. Importantly, there should be no particular bias toward closure-themed content beyond psychological factors like people making peace because culture/religion encourages it (which is not a brain imperative but a social one). If sedation is heavy, a purely predictive model might actually predict a flat affect (neutral not by balance but by numbness). And any terminal lucidity or "final rally" events would be seen as random neural fluctuations or medical phenomena, not systematically timed for reconciliation.

Fail condition: If, after adjusting for all the known variables (medication levels, presence of family, etc.), we observe that some individuals' end-of-life affect actually becomes more volatile, or remains strongly skewed (e.g. some people die in extreme despair or euphoria with no reduction in intensity), that challenges LoF. Similarly, if content analysis of final conversations and behaviors shows no enrichment of closure themes (i.e., it's just as likely for a dying person to talk about mundane or even new future-oriented things as it is to wrap up life stories), then the supposed fairness constraint isn't manifesting. Essentially, if end-of-life experiences look like a simple function of care practices and personality with no common push toward neutrality, LoF loses credence.

15.4.9 Sequence stall under short horizons (policy-level behavior)

Design: A sequential decision-making task (or multi-step game) where some policies involve delaying repair and taking risks for big rewards, while other policies allow for near-term closure at the cost of smaller rewards. For example, imagine a game with 5 steps: one strategy ("high Φ " policy) lets you periodically bank your earnings or resolve sub-goals (so you can't lose too much, ensuring partial closures along the way), whereas another strategy ("low Φ " policy) yields a higher payoff if you complete all 5 steps perfectly but offers no chance to mitigate losses if things start to go wrong. We manipulate the horizon by sometimes cutting the sequence short unexpectedly or telling the participant upfront they only have, say, 2 rounds left vs. many more. We measure at what step people abort or deviate from an ongoing plan, and neural signals of conflict if possible (maybe EEG or fNIRS if in lab).

LoF prediction: When the horizon is short (either actually or subjectively), sequences that have low Φ (meaning they require enduring risk or pain until the very end with no interim fixes) will tend to stall or be abandoned more often, even if their theoretical reward is higher. We'd expect participants to prematurely terminate or switch away from a low- Φ policy as they sense time is running out – effectively, an unwillingness to continue a plan that has no opportunity for closure when the end is near. Physiologically, ACC might show

rising activity (cost signal) as a low- Φ sequence progresses under a short horizon, reflecting mounting internal pressure to stop. By contrast, high- Φ sequences (with frequent small closures) should see steadier completion rates; people stick with those even under time pressure since they keep the ledger tidy at each step. These differences should persist after controlling for ordinary factors like cognitive load or difficulty (we'd ensure both types of sequences are equally challenging in terms of steps or memory).

Rival prediction: A non-LoF view would attribute any “stalling” simply to cognitive load or time pressure without regard to repairability. If a sequence is too complex or time is almost up, people might quit – but not selectively based on whether the plan had opportunities to correct itself. In other words, if we design the task such that both strategies are equally difficult and rewarding, a standard model might predict no difference in completion except that short horizons make people rush or possibly quit both strategies equally when the clock is nearly done. No special structure (like the presence of sub-goal closures) should systematically affect drop-out rates unless it confers some obvious immediate reward.

Fail condition: If carefully controlled experiments show that participants do not show any special aversion to low- Φ sequences as the horizon shrinks – for instance, they continue an all-or-nothing risky task to the bitter end just as often under short horizon as long, provided the expected value is the same – then LoF’s idea of an intrinsic “policy bias” toward closure is incorrect. It would suggest people only care about immediate rewards and known risks, not the reparability of a plan. That would be a failure for the LoF prediction.

15.4.10 Social menu coupling (multi-agent constraint)

Design: A multi-player network experiment or simulation where agents are connected and can affect each other’s choice sets. For example, a game where each player has a set of options, but some options remove or change options available to others (think of a shared resource or a scenario where one person’s decision closes a door for someone else). We independently manipulate each agent’s “ledger state” (e.g., give some a series of losses to put them in negative affect, others neutral or positive) and also their perceived horizon (maybe some are told this is the final round for them while others expect more rounds). We observe how often agents help or hinder each other, how options propagate or vanish through the network, etc.

LoF prediction: We should see coupled admissible sets in the group: when many agents are in a ledger-negative state (i.e., a lot of people “need” good outcomes) and horizons are short, the system as a whole will tend to redistribute opportunities to ensure

someone gets relief. For instance, help-seeking and help-giving behaviors will co-vary with horizon cues: an agent who's running out of time and deeply negative might prompt neighboring agents (even at some cost to themselves) to provide an option that helps them close their ledger (like sharing a resource, or choosing an option that opens a new opportunity for the struggling agent). Overall, LoF predicts a kind of network-level compensatory redistribution: options get reintroduced or shifted around the network so that not everyone ends up negative at the end. This might manifest as increased cooperation or an almost "rule"-like pattern where if too many are down, some spontaneously adopt an altruistic role to balance things. The key is that this happens in response to collective ledger states and horizon, not just tit-for-tat or reputation.

Rival prediction: Standard multi-agent models (like in economics or game theory) can produce helping behavior through mechanisms like reciprocity, kin selection, or reputation gain ("I help you now, someone helps me later") – but they wouldn't predict a systematic horizon-based redistribution purely to ensure fairness. In fact, if everyone is about to die (no future interactions), standard theory might predict less cooperation (no future to benefit from reciprocation). So if we remove obvious incentives, a classical view says agents should either all selfishly scramble or just stick to their learned policies. Any helping that does occur could be explained by known factors (some people are altruistic by trait, or it was in their interest via reputation in earlier rounds). There should be no mysterious tendency for the group to ensure everyone's ledgers approach neutral at the end.

Fail condition: If after accounting for reciprocity and social norms we find no residual pattern of "the network rescuing its members," then LoF's multi-agent extension is unsupported. For example, if those who are ahead or doing fine do not systematically alter their behavior when others are in dire states and time is short (beyond normal sympathy or self-interest), then there's no hidden fairness law operating at group level. A stark failure would be if, contrary to LoF, groups consistently allow or even cause some members to end in heavy deficit while others end positive, without any emergent balancing behavior even in the final rounds.

15.4.11 Blind protocols, preregistration, and negative controls

To ensure these tests are credible, we build in rigorous controls:

Blinds: We use independent teams for different stages of each study. For example, one team designs stimuli, another collects data, another codes qualitative data (like dream reports or hospice narratives) without knowing which condition is which, and yet another does the analysis. Dream content and end-of-life transcripts are double-blind coded

against preregistered dictionaries (for keywords like “forgive,” “reconcile,” “help,” etc. under a category of closure themes). This prevents subtle biases from coloring the results.

Preregistration: All hypotheses, experimental protocols, and model analysis plans are preregistered (e.g. on the Open Science Framework) before data collection. In particular, the critical interactions (like $\text{Horizon} \times \Phi$ in behavior and neural contrasts) and specific ROI predictions are declared in advance. This guards against data mining and ensures that if we find something, it was predicted, not cherry-picked.

Negative controls: We include sham conditions to verify that we’re seeing specific effects, not general artifacts. For instance:

- **Horizon shams:** We might have some framing that is superficially ominous (urgent-sounding instructions) but doesn’t actually change the person’s measured sense of future time perspective, to ensure that any behavior change is due to actual horizon shift and not just stress.
- **Content shams:** In dream studies, we could score some irrelevant categories (e.g. how much color or motion is described in the dream) to show that only the targeted repair-related content shows the predicted differences, whereas “control” content does not systematically vary.
- **ROI shams:** In neural data, we’d look at some control regions (say primary visual cortex or primary motor cortex) where we should not see any Φ -related signals. If our analyses falsely showed effects everywhere, it would indicate a non-specific result. We expect, say, no spurious ledger-related residuals in V1 or M1.

Positive controls: We also replicate known effects within our studies to confirm that the setups are working. For example, in the volatility task, we should see the established correlation between environmental volatility and pupil dilation (a known coupling). In the horizon task, we might include a condition known to induce urgency and see if people indeed respond faster or more impulsively – just to validate that participants are paying attention and behaving sensibly. If our experiment can’t even reproduce a well-known result, we wouldn’t trust it on novel predictions.

15.4.12 Statistical signatures and stopping rules

For analysis, we focus on detecting the interaction effects that would signal LoF and making sure we’re not fooling ourselves with cherry-picking or optional stopping:

Primary statistics: We expect cross-level interactions in mixed-effects models to be the smoking gun for many of these tests. For example, in the behavioral horizon experiment,

a simplified mixed model might be: $\text{Choice} \sim \text{Utility} + \text{Risk} + \text{Volatility} + \text{Horizon} + \Phi + \text{Horizon} \times \Phi + (1 + \text{Horizon} \times \Phi | \text{Person})$. Here, a significant $\text{Horizon} \times \Phi$ interaction term (with a nonzero random-slope distribution) would support LoF. In neuroimaging, we'd include parametric modulators for Φ and λ_t within our ROIs and look for significant effects surviving cluster correction or small-volume correction. Essentially, our stats are set up to find that “something extra” – a systematic effect at the combination of horizon and compensability.

Sequential analysis: We plan interim analyses with alpha-spending protocols (group-sequential design) if data collection is lengthy, to avoid p-hacking via optional stopping. That is, we set checkpoints where we can peek at data with very stringent thresholds and only stop early if effects are already overwhelmingly clear; otherwise we continue to the full sample. This prevents us unconsciously stopping when we see a hint of the desired effect.

Robustness: We report things like S-values (surprise indices, essentially $-\log_{\text{base } 2} p$ -values: “how many bits of surprise in this result?”) and Bayes factors alongside traditional p-values, to convey strength of evidence beyond arbitrary cut-offs. We also perform multiverse analyses – analyzing the data under a range of reasonable preprocessing and modeling decisions – to ensure the findings aren't an artifact of one particular analysis path. If an effect only shows up with one particular data filter or scoring method, we don't trust it. We also compare models using out-of-sample metrics like WAIC (Widely Applicable Information Criterion), LOO (leave-one-out cross-validation), and predictive log-loss to ensure that an LoF-augmented model actually generalizes better than a rival model. This way, we aren't just fitting noise – we're requiring that including the LoF constructs improves predictive accuracy on new data.

15.4.13 What would count as strong evidence for LoF

We would consider LoF strongly supported if we saw convergent results across multiple fronts. For example:

- Convergent Φ -residuals in behavior and neural data: If our experiments show that people's choices and brain activity include a significant Φ component that scales with horizon, and this holds up after aggressive controls for known factors, that's a big win for LoF.
- Dream counterweight inversion replicated: If multiple labs (perhaps different cultural settings too) find that dream valence reliably flips opposite to daytime imbalance, with blinded scoring and proper controls, this would indicate an intrinsic compensatory mechanism at work.

- Terminal variance compression with closure content: If hospice studies reveal that even after accounting for medication, patients exhibit an end-of-life affect pattern consistent with LoF (reduced variance plus meaningful last acts) across different care settings, that's powerful evidence of a general law-like constraint.
- Network-level coupling beyond known mechanisms: If in group experiments we observe a redistribution of opportunities (or a spontaneous helping network) that can't be explained by standard reciprocity or reputation models, that suggests there's a fairness constraint emergent in multi-agent systems too.

If we get any two of these four findings robustly, it would already put substantial weight behind the idea that a new constraint-law (LoF) is operating. If all four came through strongly, it would make a compelling case that, beyond doubt, something beyond standard models is needed – essentially confirming LoF's existence within empirical accuracy.

15.4.14 What would count as strong evidence against LoF

On the flip side, LoF would be on very shaky ground if we saw a consistent pattern of null or contrary results on its distinctive predictions:

- Null Horizon $\times \Phi$ interactions across multiple well-powered, preregistered studies. If behavior and neural data show no hint that short horizons amplify a bias toward compensatory choices or signals (beyond ordinary urgency or risk aversion), then the core behavioral prediction fails.
- No QS residuals in key brain regions. If careful modeling finds that vmPFC, ACC, rIFG, insula, etc. have no extra variance linked to a fairness or closure variable once we account for utilities and uncertainties, then the neural evidence for any fairness mechanism is absent.
- Dreams track salience, not counterbalance: If in the sleep studies we find that dreams simply reflect what was most emotionally salient (a bad day gives bad dreams, a good day good dreams) with no compensatory inversion on average, then the claim that dreams help balance the ledger is unsupported.
- End-of-life dispersion increases or stays high: If data showed that some people's affect variance actually increases as they approach death (or just doesn't compress) after controlling for care, or that there is no uptick in closure content compared to any random time in life, then LoF's endgame prediction fails.
- Menu coupling explained by known mechanisms: If our social network experiments find that any helpful coupling can be fully explained by things like

individuals acting on prior agreements or simple reciprocity (and if removing those factors leaves no residual balancing behavior), then there's no new law-like social effect.

If these outcomes piled up, Occam's razor would favor dropping LoF. Under such a scenario, LoF would have to retreat to perhaps a metaphysical or poetic idea rather than a scientific law, or be abandoned entirely. Essentially, we'd conclude that all the patterns we thought hinted at a fairness constraint are better explained by existing theories (adaptation, homeostasis, etc. with maybe some biases and culture thrown in).

15.4.15 Practical implementation notes

To actually do these studies right, a few final considerations:

Power: We aim for large sample sizes, given some of these effects might be subtle. For behavioral and fMRI studies like 15.4.5–15.4.6, we'd target $N \geq 120$ participants (within-subject designs give us more power, but we still want lots of data per cell). For the sleep/dream study, which is intensive, maybe $N \geq 40$ individuals with 2 weeks of nights each – a lot of data points across nights. For the hospice observations, $N \geq 100$ patients, recognizing attrition and variability will be high, and we need enough to do subgroup analyses (different illnesses, cultures, etc.).

Open materials: Everything – tasks, code, dream coding dictionaries, even synthetic datasets simulating our models – will be shared openly. This is critical since LoF is a bold claim; independent teams must be able to replicate or probe our analyses. We'll also provide pre-registered analysis notebooks and ensure that if any calibration or anchor changes (say we adjust how HCI is computed in a later revision), it's documented and new data can be labeled accordingly.

Ethics: Especially for the end-of-life and dream interventions, ethics are paramount. We will have independent ethical monitors for hospice studies (to ensure patients and families are comfortable, and to stop data collection if it interferes with care). We explicitly avoid deception in end-of-life contexts – participants and families deserve full transparency. And we commit to not inducing extreme suffering in any experiment (even the “negative days” in the dream study will use moderate stressors within ethical limits, nothing traumatic). In all cases, participants maintain the right to withdraw at any time.

Calibration: Before comparing different groups or time points, we must ensure our measures are measurement invariant. For instance, the Hedonic Composite Index (HCI) we use to quantify affect should have configural and metric invariance across different demographic groups, cultures, or even states of consciousness. We will test that a one-factor model of affect holds the same meaning across groups (configural) and that the

scale of that factor is comparable (metric) before pooling data. This prevents us from mistaking group differences in how people use the scale for actual differences in affect.

Design control: In behavioral tasks, we enforce a minimum menu-size dispersion – basically, ensure that choice sets are rich and varied enough that any bias isn't an artifact of a trivial or forced choice. For example, if a menu only has two nearly identical options, “tilt” can't be observed. We design tasks so that admissible sets have a meaningful spread in Φ and other attributes. This avoids scenarios where a horizon effect might be masked or faked by ceiling/floor effects in choice options. All experiments also include attention checks or comprehension checks to confirm participants understand the setup (especially important for framing horizon and reparability correctly).

15.4.16 Where we go next:

Bridge to 15.5: The independence tests we've outlined are only valuable if we truly follow the evidence wherever it leads. In the next section, we ask: If the rival theories (predictive coding, homeostasis, etc.) end up explaining all these results better than LoF does, what then? We sketch how the Law of Fairness research program should respond in that scenario – what it would learn and how it would redirect its efforts if the data say “no law needed.”

15.5 If Rivals Win, What LoF Learns

If the rival frameworks can explain behavior and neural data without invoking a neutrality constraint, LoF should step down from a claimed “governing law of nature” to something like a long-horizon limit of adaptive control. In other words, perhaps given enough time and intact neural mechanisms, agents tend toward neutrality simply because they minimize chronic prediction error, avoid unbounded costs (no organism tolerates infinite pain or pleasure without adaptation kicking in), and exploit social buffering. That reframes the claim as asymptotic and conditional rather than universal and guaranteed. LoF would become “a tendency that emerges in ideal conditions” instead of a strict law.

15.5.1 Narrow LoF’s domain of validity

We would refine where LoF might still hold. Instead of “all sentient beings, always,” we’d focus on niches where standard theories struggle but LoF might still operate. For instance:

- Late-life situations where consciousness is intact but horizons are objectively shrinking (perhaps LoF is only evident in the dying phase of life under certain conditions).
- Severe adversity contexts where organisms have structured opportunities for repair (maybe LoF manifests in how communities respond to collective trauma, etc.).
- Highly coupled social networks where individuals’ outcomes are interdependent (maybe fairness dynamics emerge only in group contexts, not in isolated individuals).

If evidence suggests neutrality signatures appear only (or much more strongly) in these specific contexts, then LoF isn’t a universal law but a contextual regularity. It might still be very important in those contexts, but we’d stop claiming it holds across the board.

15.5.2 Keep the measurement wins

Even if the big theory fails, the tools we developed remain valuable. The composite affect indices (HCI and the idea of Hedonic Composite Units), the dream coding techniques, the horizon-manipulation tasks, the rigorous invariance testing, the use of WAIC/LOO in model comparisons – all of these enrich the methodological toolkit of affective science. These innovations can improve how we measure well-being, how we design experiments in psychology and neuroscience, and how we account for cultural differences in emotional reporting. In fields like psychiatry, palliative care, or even AI modeling of

emotion, these tools could advance the state of the art regardless of LoF. In short, we bank the progress in psychometrics and methods.

15.5.3 Update QS from regulator to heuristic

Recall the Queue System (QS) we posited as the brain's mechanism for regulating admissible options in service of fairness. If no Φ -related residuals survive scrutiny, we would reinterpret QS not as an innate fairness governor but as a learned meta-policy. That means the brain's control hubs (ACC, rIFG, etc.) might still prune options and enforce some order, but only using known signals like utility, risk, uncertainty, and social reward – not a built-in fairness mandate. Any “fairness-like” behavior (like choosing to apologize or wrap up loose ends) would then be seen as the result of normal learning processes (perhaps taught by culture or experience) rather than an intrinsic drive. QS as a concept wouldn't vanish, but it would predict when menus shrink (e.g. under fatigue, stress, or uncertainty) without claiming they shrink to guarantee neutrality. We'd acknowledge QS as a useful heuristic description of control under constraints, not evidence of a cosmic balance law.

15.5.4 Absorb rivals' strengths explicitly

We would actively incorporate the lessons from whichever rival theory prevailed:

- From predictive coding/FEP: We'd fully embrace precision-weighted prediction error as the explanation for affect. We might say the “ledger” was always metaphorical – it really refers to expected future prediction errors. We'd integrate interoceptive inference deeply: maybe the reason we thought of a ledger is just that the brain tracks cumulative interoceptive prediction error. In essence, we'd treat a lot of LoF's language as colorful description of predictive regulation.
- From RL and homeostasis: We'd put front and center the idea of set-point regulation, allostasis, and reward maximization. For example, maybe the reason lives seem to balance is because organisms learn to avoid situations that are too extremely bad (or unsustainably good) – not due to a law, but simple survival optimization. We'd explore reformulating LoF phenomena as emergent outcomes of RL agents with risk aversion and social interconnectedness. Perhaps what we called “compensatory acts” are just normative behaviors reinforced over evolutionary time (help others now so they help you later, etc.). Essentially, we'd see if LoF patterns can be recreated by mixing predictive coding with RL and homeostatic drives, without needing a new principle.

15.5.5 Salvage the philosophical core without overreach

One thing we likely wouldn't abandon is the moral aspiration that motivated LoF: the idea that a just world would ensure no life is overwhelmingly positive while another is overwhelmingly negative without eventual balance. Even if nature doesn't guarantee that, it's a powerful ethical concept. We might pivot to saying: "Okay, there's no physical law of fairness – but maybe our societies should strive to create one." In practical terms, translate LoF into design goals for social and health policy: ensure every person has access to "repair" mechanisms (mental health care, opportunities for redemption), build institutions that allow late-life reconciliation (hospice programs that include family healing sessions, for example), use what we learned about dreams and coping to improve well-being (maybe encourage practices that induce compensatory dreaming or reflection). We would frame LoF not as a fact of nature but as an ideal to pursue: for instance, argue that the healthcare system should aim not only to keep people alive (viability) but to help "close their ledgers" (through pain management, psychological support, etc.) by life's end. The absence of a natural law doesn't make the idea worthless – it just means it's up to us (conscious agents and societies) to implement fairness, rather than expecting physics to do it for us.

15.5.6 Redirect experiments, not rhetoric

We commit to transparently publishing any null or disconfirming results – no sweeping under the rug. All code and data would be out there for others to analyze. We'd then reallocate our research efforts to new questions that arise from LoF's failure. For example: What minimal ingredients did give us those tantalizing near-successes? Maybe our work showed that adding a certain social component to RL algorithms produces quasi-fair outcomes – that's worth studying in AI ethics or multi-agent simulations. Or, which clinical interventions (perhaps unrelated to an innate law) most effectively produce the appearance of fairness? Maybe we find that certain hospice practices lead to peaceful death (neutral affect) more often – not because of a law, but because those practices work; we'd double down on those findings for practical benefit. We'd also ask, how do cultural norms manufacture perceived neutrality? Perhaps rituals of apology, forgiveness, or confession exist precisely because there is no natural guarantee – cultures might have evolved them to force a kind of fairness. That's an anthropological and psychological question we'd pursue: how humans actively create balance through narrative, religion, justice systems, etc., and how effective those are. The point is, even if LoF as a natural law dies, the research program can pivot to understanding how approximate fairness happens and how we can encourage it.

15.5.7 Define a clean retirement rule

We won't keep LoF on life support indefinitely. We'd establish clear criteria for declaring the law "retired" as a scientific hypothesis. For instance: if after, say, three multi-lab, pre-registered replications of each key independence test we consistently see (i) no Horizon $\times \Phi$ interaction in behavior, (ii) no QS-like residuals in the neural signatures in ACC/vmPFC/rIFG/insula, and (iii) no evidence of dream counterweights or terminal variance compression beyond conventional explanations – then it's time to call it. At that point, we formally publish that LoF did not hold up and should not be considered an active scientific model. We'd encourage the community to take whatever useful parts remain (as metaphors or tools) and otherwise move on. In essence, we'd issue a "death certificate" for the law claim, to avoid zombie theories lingering just because of sunk cost or emotional investment. We keep the measurement toolkit and any insights gleaned, but we release the grand claim.

15.5.8 Document the epistemic gain

Even a failed big hypothesis can leave science stronger. We will catalog the improvements and insights gained in the process of challenging LoF. For example, maybe the effort to test LoF forced us to develop a truly robust cross-cultural mood measure – that's a win for psychology. Or perhaps it brought together hospice workers and neuroscientists in a new way, yielding protocols that will be useful in palliative care research. Maybe it set a new standard for adversarial collaboration, where proponents of different theories worked together – that's a template for future hard problems. We'll emphasize these contributions in our concluding reports: that trying to chase LoF forced us to ask questions that were previously glossed over (like "what is the combined distribution of lifetime happiness across people, and why?") and to create infrastructure (data, code, collaborations) that now can be redirected to other questions. In short, even if LoF is wrong, the pursuit of it will have produced lasting knowledge and tools.

15.5.9 Leave a small door open – properly

Good science is open-minded but not wishy-washy. If LoF mostly fails but one narrow anomaly remains – say, for example, multiple studies find that even after controlling everything, there is still a little unexplained reduction in affect variance in the final days of life that we can't attribute to meds or psychology – we won't just shout "Eureka, the law lives!" Instead, we'll isolate that anomaly as a specific puzzle: maybe turn it into its own hypothesis ("there is a terminal dampening effect of unknown origin") and design new experiments focusing only on that. We would not use a leftover anomaly to grandly resurrect LoF without solid evidence. Instead, we keep that door slightly open for further

inquiry, while being clear that LoF as originally stated is (at best) extremely limited or conditional. If future evidence later revises that, fine – but it must meet the same high bar we set.

15.5.10 Bottom line

If the rivals win, LoF becomes a useful scaffolding that helped us build better science, but not a law of nature. We will have advanced measurement and sharpened questions, and we'll have a clearer view of human affect dynamics even in LoF's absence. The research program doesn't die – it evolves, becoming leaner and humbler, still guided by the same human concern (why does suffering feel distributed the way it is, and how can it be managed?), but now firmly led by data rather than by the initial grand idea. In the end, whether LoF stands or falls, we prioritize truth over tenure: the narrative must be decided by evidence, not by our attachment to a theory.

15.5.11 Where we go next:

We now move to Chapter 16 – Reinforcement Learning and Homeostasis, which builds the strongest possible case for these mainstream frameworks as rivals to LoF. In that chapter, we clarify how standard reward optimization and homeostatic regulation drive behavior and maintain stability, highlight why optimizing for rewards or set-points isn't the same as ensuring a neutral lifetime ledger, and lay out where these mechanisms overlap with LoF's predictions versus where LoF demands something extra. Chapter 16 sets the stage for empirical tests that will ultimately let data decide between smart regulatory control and a true fairness constraint.

Chapter 16 — Reinforcement Learning and Homeostasis

Most contemporary brain theories explain behavior without invoking any cosmic notion of fairness. They start local and mechanistic: organisms optimize expected reward under constraints, and they stabilize internal variables like temperature, glucose, and arousal. Two major families dominate this story:

Reinforcement learning (RL) – The brain updates its policies by comparing what it expected to happen with what actually happened. A good surprise nudges you to repeat an action; a bad surprise nudges you to change course. Phasic dopamine signals, striatal plasticity, and cortical value maps are the biological footprints of this process. Over time, an agent learns what to do in which states to maximize cumulative return.

Homeostasis (and allostasis) – Life survives by keeping key internal variables within viable ranges. Classic homeostasis defends fixed set-points (like a thermostat); allostasis goes further, predictively adjusting those targets based on context and anticipated demand. In the brain, this looks like layered control loops that regulate autonomic tone, endocrine release, sleep pressure, and effort—often before conscious awareness catches up.

Together, RL and homeostatic control can produce a world that looks orderly without any appeal to fairness. Habits form; cravings ebb; pain teaches avoidance; rest restores function; social bonds are reinforced because they reduce risk and increase long-term reward. Given enough time and supportive conditions, such a system can appear to “balance out” distress with relief simply because grossly imbalanced organisms function poorly and will learn to avoid ruinous extremes.

This chapter makes that rival picture as strong as possible, then shows where the Law of Fairness (LoF) asks for something more. We will: (1) clarify the core commitments of RL and homeostatic control (what must be true for these mechanisms to explain behavior); (2) distinguish optimization from balance — why maximizing returns or minimizing deviations does not automatically guarantee a neutral lifetime ledger of felt experience; (3) locate the overlaps where RL/allostasis and LoF predict similar outcomes (recovery after loss, the value of sleep, the late-project surge in making amends); (4) mark the divergences — signatures LoF predicts that optimization alone does not (e.g. horizon-sensitive pruning of options even when immediate rewards are equal; end-of-life variance compression beyond medical symptom control; dream counterweights after unusually hard days); and (5) specify empirical tests to cleanly separate “smart control” from “fairness enforcement,” so that data – not philosophical preference – decide between them.

We proceed in three broad steps. Section 16.1 unpacks rewards, set-points, and the power of caregiving via everyday examples (the athlete tapering before a race, the new parent re-weighting sleep vs. work, the anxious patient who finds relief after one clear diagnosis). Section 16.2 then shows why optimization isn't balance: an agent can rationally choose a high-variance, high-reward life path that leaves large hedonic debts, and nothing in vanilla RL or allostasis ensures the lifetime ledger will wash out. Section 16.3 presents composite rivals ("H*" hybrids) that blend RL, homeostatic, predictive coding, and social mechanisms — asking whether a clever emergent mix could mimic LoF's signatures without invoking any fundamental fairness constraint. Next, Section 16.4 provides research notes on how to fairly compare models and perform adversarial fits (ensuring LoF doesn't "win" just by beating a strawman), and Section 16.5 outlines what the LoF framework would keep or change if RL-homeostasis (or a hybrid) ends up outperforming LoF in head-to-head tests.

For readers, the practical upshot is this: RL and homeostasis already explain a lot of what life feels like — habits, cravings, fatigue, "second winds," the quiet relief after closure. The Law of Fairness enters only where those mechanisms, taken to their logical end, still fail to guarantee that the lifetime ledger of feeling ends neutral. In other words, LoF posits an extra constraint (a fairness "guardrail") operating above ordinary optimization. This chapter is about deciding whether that extra guarantee is real or whether smart regulatory control was sufficient all along.

What you'll get from this Chapter:

- Core ideas of RL and homeostasis clarified: A solid understanding of how reinforcement learning and homeostatic regulation drive behavior and maintain stability, illustrated with relatable examples (from athletes tapering to the body's anticipatory adjustments).
- Why optimization isn't automatically fairness: Insight into why simply maximizing rewards or keeping variables in range doesn't guarantee a balanced life ledger. You'll see how an agent could accumulate a large "hedonic debt" even while behaving "rationally" by standard criteria.
- Where control mechanisms mimic fairness: Identification of scenarios where ordinary RL/allostasis produce LoF-like patterns (e.g. recovery after setbacks, the value of rest and late-in-life reconciliations) and why these emerge naturally even without a fairness law.
- Telltale signs beyond standard models: An inventory of specific signatures that pure optimization can't explain – such as choices narrowing as horizons shrink,

end-of-life emotional variance compressing beyond medical causes, or dreams flipping in tone after hard days – which point to an additional fairness constraint at work.

- Data deciding the debate: A look at how experiments will pit “smart control” against “fairness enforcement.” You’ll learn what evidence could confirm that LoF’s extra guarantee is real (for example, if models including a fairness term consistently predict outcomes better by an information criterion like WAIC) and how we ensure our tests are unbiased (adjusting for any overdispersion in count data by using Negative Binomial models when needed).

Subsections in this Chapter:

- **16.1 Rewards, Set-Points, and Care** - Uses everyday scenarios (the competitive athlete, the sleep-deprived new parent, the anxious patient finding relief) to illustrate how reinforcement learning and homeostatic adjustments operate.
- **16.2 Optimization Isn’t Balance** - Demonstrates that an agent could rationally choose high-reward, high-variance paths that leave an imbalance of pleasure and pain. This section explains why nothing in vanilla RL or allostasis ensures that the lifetime hedonic ledger will wash out to zero.
- **16.3 Composite ‘H’ Hybrids** - Introduces blended models that combine RL, homeostatic, predictive coding, and social mechanisms. We ask whether an emergent mix of standard processes could imitate all of LoF’s signatures without assuming any fundamental fairness drive.
- **16.4 Research Notes – Fair Model Comparisons** - Outlines how we rigorously compare LoF with rival models on equal terms. This includes using a single out-of-sample metric (like WAIC) to judge predictive fit and checking statistical assumptions (e.g. switching from Poisson to Negative Binomial if count data show variance/mean > 1.2), so that LoF doesn’t win by default due to biased methods.
- **16.5 If a Rival Wins:** Discusses how the LoF framework would adapt or what it would relinquish if RL-homeostasis (or a hybrid model) outperformed LoF in direct tests. We consider which aspects of fairness might still hold and which would need rethinking if a simpler “smart control” theory fits the data better.

Where we go next:

We begin with ground rules for RL and regulation—what “reward,” “set point,” and “care” mean operationally—then test where optimization diverges from balance. Section 16.1 starts by defining rewards, set points, and the roles of caregivers as measurable variables rather than motives. Examining these familiar situations, we set the stage for probing where these mechanisms succeed – and where they may fall short – in delivering a fair life balance.

16.1 Rewards, Set Points, and Care

If you put a thermometer and a tip jar inside a brain, you would already understand half of its daily business. The thermometer represents homeostasis: sensors, set-points, and controllers that keep internal variables within viable ranges. In fact, modern RL theory explicitly ties rewards to physiological homeostasis. Keramati and Gutkin (2014) note that “our survival depends on our ability to maintain internal states (e.g. temperature, glucose) within narrowly defined ranges”. They formalize behavior as rewarding if it reduces the distance from these setpoints. This highlights that LoF’s view (organisms strive to correct deficits) is consistent with models where homeostatic stability itself drives learning. The tip jar represents reinforcement learning (RL): a ledger of what “paid off” last time, so the system is more likely to try that action again. Most of life’s basic stability—and much of its learning—can be sketched with those two props. In this section we make that sketch precise enough to respect neuroscience, rich enough to capture everyday experience, and clear enough to separate what these mechanisms do from what the Law of Fairness (LoF) might add on top.

16.1.1 The thermostat, upgraded: homeostasis and allostasis

Classic homeostasis keeps bodily variables close to target levels—temperature, pH, blood glucose, osmolarity, etc. In the brain, these control loops run through the hypothalamus, brainstem autonomic nuclei, and endocrine partners. But organisms do more than defend a static target; they also anticipate. That anticipation—called allostasis—shifts set-points and control gains based on time of day, context, and expected demand (Ramsay and Woods, 2014). Before a sprint, heart rate rises; before sleep, core body temperature drops. Cortisol ramps up before you wake not because you were stressed during the night, but because you’re about to need glucose and alertness upon waking.

A useful mental model is to imagine nested control loops of differing speeds:

Fast loops (milliseconds to seconds): e.g. baroreflex for blood pressure, pupillary reflex, vestibulo-ocular reflexes.

Intermediate loops (seconds to hours): e.g. ghrelin/leptin cycles for hunger, insulin release, thermoregulation, inflammatory tone.

Slow loops (days to seasons): e.g. circadian rhythm phase, reproductive hormone cycles, muscle anabolism, longer-term immune adaptations.

Subjectively, you experience these loops as fluctuations in energy levels, drive, and the ease of action. When the loops are aligned and in healthy ranges, small tasks feel “light”

and rewards land with normal satisfaction. When the loops are strained or out of tune, everything feels heavier (high effort cost) and pleasures are blunted.

16.1.2 The tip jar: reinforcement learning in the wild

RL starts with a simple premise: compare what you expected to happen with what actually happened, then update your policy accordingly. In the brain, phasic bursts of dopamine from midbrain nuclei (VTA/SNC) broadcast a prediction error signal that modifies corticostriatal synapses. A positive surprise (better outcome than expected) strengthens the synapses active just before the reward; a negative surprise weakens them. Over time, the agent learns a mapping from states to actions that maximizes cumulative return (total expected reward over time).

Three practical RL features matter for lived experience:

Temporal difference learning: The system can learn from partial progress and delayed payoffs, not just immediate gratification. (It propagates rewards backwards in time to earlier cues.)

Generalization: Cortical value maps allow learned value to spread to similar states or actions. In effect, skills and preferences transfer to new but related contexts.

Hierarchies: Repeated actions become “chunked” into habits and procedures, freeing up attention and making behavior more efficient (habit learning builds subroutines).

Subjectively, these features underlie the feeling of skill and taste. You reach for the frying pan in just the right way without thinking (an efficient habit); you choose the route you “like” because past experience wired in a preference; you revisit the café where a kindness once happened because the positive prediction error left a lasting imprint. None of this requires a story about cosmic fairness or eventual balance—it only requires a nervous system that updates based on what’s useful for future outcomes.

16.1.3 When the thermometer meets the tip jar

Most of the time, the homeostatic loops talk to the RL tip jar. Keramati & Gutkin’s model also shows that *time preference* emerges naturally from homeostasis: behaviors that close deficits sooner are rewarded more. In fact, under their formal assumptions, they derive delay discounting – valuing immediate rewards over future ones – as a policy that supports faster homeostatic recovery. Adding this reinforces LoF’s idea that, as life’s end nears, people will naturally favor payoffs that balance their ledger more quickly. If you are sleep-deprived, the value of a nap goes way up; if you have just eaten, the value of dessert drops; after social exclusion, the value of a friendly voice or hug spikes. This state-dependent shift in pleasure is known as *alliesthesia* (Cabanac, 1971): stimuli that

restore homeostatic balance (food when hungry, water when thirsty, rest when tired, social contact when lonely) feel disproportionately rewarding, whereas the same stimuli provide little pleasure — even aversion — when one's needs are already satisfied. Alliesthesia exemplifies an innate balancing drive: the internal state modulates what feels good or bad so that organisms seek what they lack and avoid what they have in excess, thereby maintaining equilibrium. We can sketch this coupling formally as a state-dependent value function:

$Q(s, a) = U(a | s) + \Delta\text{Homeo}(s, a) + \gamma E[V(s') | s, a]$, where $Q(s, a)$ is the overall value of action a in state s . Here, $U(a | s)$ captures the straightforward immediate reward or relief from action a ; $\Delta\text{Homeo}(s, a)$ captures how the action nudges bodily or psychological variables toward or away from their viable ranges (a homeostatic “bonus” or “penalty”); and γ is the usual discount factor for future rewards $V(s')$. Actions that repair depleted resources or reduce internal strain (sleep, hydration, warmth, reassurance) earn an extra bonus via ΔHomeo because they improve the very platform on which all other future rewards depend. This structure resembles the “homeostatic reinforcement learning” framework (Keramati and Gutkin, 2014), which explicitly integrates reward maximization with physiological stability goals. Three everyday examples: to illustrate this RL-homeostasis coupling, consider:

The athlete’s taper: In the week before a major race, the athlete dramatically reduces training volume. Allostastic loops seize the opportunity to reduce inflammation and restore glycogen; meanwhile, RL has learned that short-term rest maximizes long-term payoff on race day. The subjective signature is restlessness with purpose: the athlete feels an urge to train (habit), but “knows better” and holds back to reap a bigger reward later.

The new parent’s reprioritization: A combination of sleep deprivation, oxytocin release, and heightened responsibility reshapes the value landscape. A two-hour nap suddenly feels far more rewarding than a leisurely dinner out. RL encodes these new payoffs (experientially learning that sleep is precious), and homeostatic sleep debt makes the reprioritization feel like a necessity rather than just a preference.

The “diagnosis effect”: Severe uncertainty (not knowing what’s wrong) amplifies stress and arousal; getting a clear diagnosis—even before any treatment—often brings relief and restored sense of control. In physiological terms, the body quiets an overactive stress response, and actions that once felt impossible (due to fatigue or anxiety) suddenly have traction. Part of the relief is informational (reducing uncertainty), and part is homeostatic (lowering allostatic load when the “unknown threat” is resolved).

16.1.4 Care as control

Caregiving is one of biology's smartest control policies. Social contact, touch, calming vocalization, and predictable presence have measurable regulatory effects: they increase heart-rate variability, reduce cortisol, and even modulate pain (through descending nociceptive inhibition). In the brain, regions like the anterior insula integrate interoceptive (body-state) prediction errors; ventromedial PFC (vmPFC) integrates social value with internal state; the anterior cingulate cortex (ACC) helps arbitrate effort and conflict. Similarly, cognitive emotion-regulation strategies (like reappraisal) deliberately engage prefrontal control networks to dampen limbic activity, thereby reducing negative affect (Ochsner & Gross, 2005). In essence, the brain can consciously coach itself back toward equilibrium by reframing how it interprets a situation. In RL-homeostatic terms, care increases the expected return of doing hard things by lowering their physiological cost. Being supported also expands the action set: when you feel safe and backed up, more options become feasible that you wouldn't even consider under threat or exhaustion.

You can think of skilled caregiving as a combination of set-point management and value shaping:

Set-point management: Stabilize the basics – ensure the person gets sleep, nutrition, pain relief, and a steady daily rhythm. By keeping homeostatic variables in healthy ranges, care minimizes “background noise” in the system.

Value shaping: Scaffold small wins, reinforce acts that provide relief or repair, and de-value impulsive self-harm by raising its perceived cost. (E.g. a caregiver might celebrate a patient's minor improvement, reinforcing it, and gently dissuade choices that would worsen their condition.)

None of this presupposes any cosmic ledger of fairness at work. It simply shows how good control architectures (biological and social) generate lives that often rebalance after strains because rebalanced systems function better. In short, even without invoking LoF, a well-regulated system with feedback, adaptation, and care will tend toward recoveries and counter-balancing behaviors.

16.1.5 What optimization explains well

Give the RL-plus-homeostasis framework strong credit where it shines. Many patterns that make life seem self-correcting can be explained mechanistically by these tools:

Habituation and hedonic adaptation: The first bite of dessert is best; repeated exposure yields lower marginal reward. Systems naturally normalize to steady inputs.

Opponent after-effects: Coffee up, crash down. A hot sauna followed by a cold plunge brings relief. Homeostatic loops often overshoot and then self-correct, producing a rhythm of opposite after-effects.

Recovery after losses: After setbacks, people often reconnect socially, catch up on sleep, or find meaning-making activities that restore functionality. Learning processes also update behavior to avoid the same harm again.

Why rest and ritual help: Sleep consolidates learning and resets homeostatic control gains; familiar rituals reduce uncertainty and rally cooperative support. These promote stability and resilience without any cosmic balancing force.

(We will revisit these everyday balancing tactics in Chapter 21, recasting habits, social support, and rituals as a personal “ledger gym” for emotional equilibrium. It’s important to note, however, that here we describe their effect without invoking any special fairness law – they are part of our brain’s normal toolkit. LoF, if real, would incorporate these tools as well, but it claims something more: a guaranteed lifetime balance beyond what ordinary coping ensures.)

If LoF vanished tomorrow, one could still tell a compelling, entirely mechanistic story for much of life using just these two tools (RL and allostasis). In fact, evolutionary biology and psychology have largely done exactly that.

16.1.6 Where optimization is silent

However, there are gaps—areas where standard optimization has no say about lifetime fairness: No lifetime guarantee: An agent can rationally choose a high-variance path—entrepreneurship, frontline crisis work, extreme caregiving—that yields a large net negative emotional ledger if the rewards never materialize or come too late. Standard RL optimizes expected returns along the way; it has no mechanism to reach back at the end of life and ensure the total integral of feeling is near zero. Homeostasis will keep you viable (alive and functional), but not necessarily emotionally balanced in the long run.

Local optimization only: Decisions are made myopically, with local information and gradients. If the world deals you repeated bad draws—illness, disaster, betrayal in uncorrelated succession—neither RL nor allostasis is obliged to produce compensatory good experiences by the time your stream of experience ends. They will try to cope and adapt in each instance; they will not guarantee a neutral final ledger. In short, these mechanisms explain functional resilience, not ensured fairness.

(These points are not criticisms of RL or homeostatic theory; they’re simply reminders of what those mechanisms do not claim to do. They explain function, not fairness.)

16.1.7 Setting up the test against LoF

The Law of Fairness agrees that RL + homeostatic control should generate a strong tendency toward affective mean-reversion in many situations—because extremely imbalanced organisms perform poorly and tend to self-correct or receive care. LoF parts ways with the standard view on one bold claim: for conscious streams, the time-integral of felt experience (the “ledger”) closes within a preregistered equivalence band around neutral by the death of mind (i.e. when an individual’s sentient life ends, their cumulative total of pleasure minus pain lies within a predefined neutrality range). If that claim is true, we should observe specific signatures that optimization alone does not require:

Horizon-sensitive menu pruning (Chapter 6; 5.2): As one’s remaining time shortens, options that enable emotional repair or relief become unusually easy to execute, whereas alternative options quietly lose traction—even when those alternatives have equal immediate utility. (In other words, near the end, the choice set narrows in a way that favors “ledger-balancing” actions.)

Variance compression near end-of-life (Chapter 11): Affective variability (extreme highs and lows) narrows as death approaches—beyond what medical symptom management alone can explain. We see a system-level tilt favoring closure, reconciliation, calm, and relief-seeking behaviors, with an unusual persistence that outstrips ordinary coping.

Dream “counterweights” after very hard days (Chapter 10): After an unusually negative day, the content of that night’s REM dreams shows a valence inversion (skewing positive) that predicts an improved mood baseline the next day—beyond the standard benefits of sleep or memory consolidation.

Crucially, these signatures are testable. Chapters 10–13 detailed how one could measure each pattern rigorously. Here in Part VII (Chapters 14–16), we are staging a fair fight between LoF and its rivals using those measurements. For now, keep our two props in view: the thermostat and the tip jar already carry us far in explaining behavior. If a law of fairness exists, it would sit atop those mechanisms—not replacing the feedback loops or RL updates, but constraining the world so that across an entire life, optimization is never the whole story.

16.1.8 Where we go next:

In the next section, we examine why even such robust self-regulation can fall short of ensuring fairness over a lifetime. Section 16.2 shows that optimization and homeostasis, while keeping an organism viable and adaptable, offer no guarantee of a neutral final ledger. This sets the stage for empirical signals that help us distinguish mere resilience from a true fairness-enforcing mechanism.

16.2 Optimization Isn't Balance

Reinforcement learning and homeostatic control make organisms competent; they do not, by themselves, make lives fair. This section uses everyday scenarios and simple math to show how optimization can succeed locally yet fail to neutralize the lifetime affective ledger. Along the way, we draw out empirical signals that can distinguish mere “smart regulation” from a true fairness-enforcement mechanism.

16.2.1 Local gradients vs. lifetime integrals

RL updates policies by following local gradients in expected reward, and homeostasis stabilizes local variables around viability ranges. The Law of Fairness, by contrast, posits a global, path-wise constraint: for a conscious stream, the time-integral of experienced valence must close zero by life's end. Local success (moment-to-moment coping and learning) says nothing about global neutrality.

Mountain trail metaphor: A hiker who always walks “uphill” (steepest ascent) can still end up on a ridge far from the summit. RL’s policy-gradient ascent finds good ridges; homeostasis keeps the hiker hydrated and alive. Neither ensures the hiker finishes at sea level. LoF, however, claims the hiker’s journey will end at sea level regardless of which ridge was taken.

Formally, standard RL maximizes an expectation: $E[\sum_t r_t]$ (the expected sum of rewards over possible futures). LoF, in contrast, asserts an almost-sure path constraint:

$\int_0^T F(t) dt \approx 0$ (within a pre-registered neutrality band) as T approaches the end of the conscious stream, where $F(t)$ is instantaneous signed affect centered at 0.

In plain terms: optimizing expected return over many hypothetical lives is not the same as guaranteeing near-neutrality in the one life you actually live. An agent can “win” in expectation yet experience a very lopsided outcome in reality.

16.2.2 Viability vs. fairness

Homeostasis aims at survival, not emotional equipoise. A body can remain perfectly viable while accumulating a long sequence of negatively-valenced states:

Chronic pain: The person’s inflammatory responses might be finely tuned, cardiovascular control intact, sleep efficiency adequate—all signs of homeostatic viability—yet the person endures years of net suffering. Nothing in homeostasis prevents a persistently negative hedonic total.

Endless caregiving: Someone might manage their stress (cortisol within normal bounds, heart-rate variability stable) and continue functioning day-to-day, but their emotional

ledger drifts ever more negative because their duties continuously outrun the rewards. If LoF holds, we expect to see counterweight processes kick in to prevent these long-term deficits: dream content shifting valence, social dynamics tilting to assist, admissible choices narrowing to force rest or relief – especially as horizons shrink (see Chapters 5–6 on the Queue System, Chapters 10–11 on dream and end-of-life phenomena). Pure optimization, on the other hand, predicts coping where possible but no guaranteed balancing. It would be perfectly content with a chronically unhappy survivor, so long as that state was locally “optimal” given constraints.

16.2.3 Exploration, exploitation, and tail risk

RL must trade off exploring new options vs. exploiting known rewards. Exploration inevitably means occasionally drawing bad luck. Rare catastrophic events (illness while traveling, a startup going bust, an accident) can dominate someone’s lifetime affect.

Consider high-variance careers: entrepreneurs, performers, frontline responders. These roles face fat-tailed outcome distributions (a few huge wins, many losses). RL can rationally choose such high-variance policies based on prior probabilities of success. But there is no mechanism in standard RL to ensure that if a particular life gets a string of bad-luck outcomes, it will be compensated later before that life ends. Put differently, RL and allostasis operate as if many lives will average out (an ensemble perspective), not to guarantee that each individual path evens out.

Non-ergodicity insight: What’s optimal in expectation over many lives is not necessarily favorable for the one life you get to live. LoF, if true, implies a path-wise compensation (each trajectory individually tends to neutral), whereas RL’s logic remains ensemble-wise (some trajectories will be net positive, some net negative, and that’s acceptable under purely statistical optimality).

Empirical discriminator: In cohorts that choose or experience high variance, LoF predicts a distinctive pattern near the endgame: affective variance compression and menu shifts that can’t be explained solely by external relief (medical, financial). For example, among risk-takers who had many bad breaks, do we see an anomalous closing-of-ranks on remaining positive opportunities (and a calming of mood volatility) as if an unseen hand is mitigating their losses? Pure optimization does not require any such end-of-life pattern—it would expect many of those lives to end in persistent deficit.

16.2.4 Distribution shift and Goodhart’s traps

Policies learned under one set of conditions often degrade when conditions change (distribution shift). Moreover, optimizing a proxy measure can backfire (Goodhart’s law).

Shift: Coping strategies that worked in youth or health may fail after a serious

diagnosis; the reward landscape re-maps. RL can eventually re-learn under the new distribution, but it does not guarantee any restoration of the lifetime integral lost during the interim.

Goodhart: A person might chase social approval or income as a proxy for well-being. This yields local rewards (praise, raises) but might erode genuine relationships and health. The local feedback is positive even as the global well-being integral turns increasingly negative. LoF makes a prediction here: under prolonged proxy-chasing or life-strategy errors, we should see the Queue System (QS) kick in. That is, the person's feasible menu of actions will subtly prune away further proxy-seeking opportunities and re-weight toward reparative or meaningful actions that enable closure. If instead the person can continue chasing the misleading proxy with undiminished ease—and thus ends life with a skewed ledger—LoF would be falsified in that case. (No hidden hand intervened to course-correct.)

16.2.5 Multi-agent coupling (your menu depends on others)

In reality, we are not independent agents picking rewards from a static environment; we are embedded in societies. Other agents' choices shape your feasible options. Restaurants run out of food; attention economies saturate; housing markets price people out. RL treats these as shifting state transitions, and homeostasis adapts to the stress but neither ensures fairness across individuals.

Competition externalities: When many people pursue the same scarce good, there will be "losers" who did everything right in a relative sense yet end up with net negative affect due to external conditions. Optimization doesn't prevent zero-sum outcomes.

Adversarial dynamics: Other agents (or algorithms) learn to exploit your learned habits think scams tailored to your past behavior or predatory apps that hijack your reward system. Your local learning may lag behind these adversarial moves, leaving you worse off despite optimizing for past conditions.

LoF implies a population-level coordination: feasible sets of actions ("admissible menus") are coupled across agents such that each conscious stream still has a route to neutrality that doesn't require violating someone else's. In concrete terms, even in very crowded, competitive environments, we should detect micro "openings" for repair and relief preferentially appearing for those who have the largest negative ledgers. Pure optimization predicts the opposite: those least advantaged would continue to have shrinking options and compounding losses in zero-sum contexts. This is a difficult test—essentially asking if there's an emergent fairness in how opportunities present in groups, beyond what any policy or institution explicitly ensures.

16.2.6 Optional stopping and the martingale intuition

A common intuition says, “If ups and downs tend to even out (mean-revert), then over a lifetime it should all come out fair in the end.” This is a fallacy. In stochastic processes, even a true martingale (a fair game with no drift in expectation) can produce highly imbalanced realized outcomes by a bounded stopping time (as formalized in the optional stopping theorem). Preservation of expected value does not guarantee that the realized path ends near zero. And many real processes are actually supermartingales for well-being (they exhibit a slight downward drift under risk and loss aversion).

In RL, once you introduce risk and discounting, the value process for felt rewards is not a martingale with respect to experienced utility. There is no mathematical guarantee of ending at zero.

Homeostatic rebounds (opponent processes) do create local mean-reverting forces (e.g. every high triggers a comedown), but the cumulative integral can still wander widely. Without an extra constraint, there is nothing binding the path to close at zero by the end.

LoF is precisely such an extra binding condition. If LoF is false, we should eventually find stable sub-populations of people whose lifetime integrals of felt experience remain far from neutral—even given plenty of time, support, and all known aids. In generous conditions (long lives, good healthcare, robust social support) we would observe some individuals ending life with strongly positive or negative totals outside the “neutral range.” Robust evidence of such uncorrected extremes would directly challenge LoF.

16.2.7 Three clean separation tests

We highlight three empirical tests that offer especially clear separation between pure optimization and LoF’s added constraints. These tests reflect hypotheses that LoF requires but that RL/homeostasis (and even other rival theories) do not:

Horizon-dependent menu tilt: LoF prediction: As someone’s future shortens, their admissible set narrows preferentially toward reparative acts (saying sorry, finding closure) and away from high-risk, low-compensation acts. H* response: Short horizons naturally increase discount rates and risk aversion in decision-making. People become more cautious (“safer” choices) as time runs out. Also, families, caretakers, and institutions tend to add support near the end (hospice care, visits from loved ones), which could bias choices toward meaningful or comforting activities. In sum, H* says the tilt happens because of generic prudence and increased help—not because of a fairness-driven admissibility rule.

Variance compression near death: LoF prediction: Regardless of a person's prior highs or lows, their affective trajectory converges to near-neutral in the final stretch (additional variance gets squeezed out). H* response: In the final weeks, clinical protocols become very aggressive in managing symptoms; heavy sedation and analgesia are common, which obviously flatten affect. Social rituals around dying (everyone speaking in calm tones, spiritual support) tend to standardize the emotional climate. So the reduction in variance could be explained by medical and social factors "leveling" everyone's experience near the end. No fundamental law needed—just end-of-life care and common human ritual.

Dream counterweights after hard days: LoF prediction: After an unusually bad day, that night's dreams carry opposite-valence material that produces a measurable upswing in next-day mood (beyond normal sleep effects). H* response: REM sleep is known to help with fear extinction and processing of negative memories. So if you had a rough day, your dreams might incorporate those themes and soften their impact (which can feel like "inversion"). And any next-day relief can be attributed to generic sleep restoration and memory reconsolidation doing its job—not a ledger-balancing force.

In short, a well-tuned H* could approximate the LoF patterns. The scientific question is whether H* can match the strength, timing, and conditional specifics of those patterns across the board. LoF sets a high bar: the effects should be strong, occur at just the right times, and scale with the person's deficit. H* might mimic them in form, but can it do so consistently without slipping in an ad hoc fairness mechanism?

16.2.8 Case studies: optimization wins yet fairness fails

It may be helpful to imagine concrete life stories where RL and homeostasis appear to do everything right, yet the outcome still feels deeply unfair. These cases illuminate what LoF would need to correct if it's real:

The dutiful striver: A person spends decades doing everything society deems "right" – they work hard, care for family, make prudent financial choices. By RL standards, their policy has been optimal relative to their values; homeostatic self-care is adequate. Then, in a short span late in life, they suffer an avalanche of bad outcomes (e.g. loss of a spouse, a cancer diagnosis, financial collapse). There is nothing in standard optimization that ensures compensatory good experiences of equal magnitude will arrive before this person's life ends. The ledger can end deeply negative despite all earlier competence and virtue.

The ascetic athlete: This person optimizes their body and training to an extreme (excellent homeostatic tuning, disciplined RL for long-term rewards). But fate deals

repeated injuries that wipe out gains and leave chronic pain. Their brain networks (resting-state, etc.) might remain efficient—nothing “broken” homeostatically—yet their affective integral trends negative because pain dominates their later years.

LoF’s claim would be that in such cases, late-life shifts should open pathways to relief, reconciliation, and meaning with unusual persistence—essentially giving these individuals special chances to balance the ledger at the 11th hour. If careful observation finds that such pathways do not open (and these lives simply end in deficit), then LoF fails precisely where it would matter most.

16.2.9 Why mere “tendency” isn’t enough

It might be tempting to retreat to a weaker claim: perhaps life has a tendency toward hedonic neutralization, but no guarantee. However, statistical tendencies do not protect individuals from worst-case outcomes. A fair coin usually lands heads ~50% over thousands of flips, but your particular sequence of flips could still be lopsided. Likewise, saying “most people balance out on average” offers zero comfort or correction for the person who doesn’t. If fairness is to be a real property of our universe, and not just a hopeful sentiment, the guarantee must be path-wise (each life individually) and horizon-aware (especially active as opportunities dwindle). Otherwise, it’s not a law—just a loose generalization.

Ethical stakes: Public health and justice systems cannot be built on “most people roughly even out.” They either have to accept that some people will end up with irredeemably unfair lives, or they have to posit a law-like closure mechanism that ensures balance for everyone. LoF takes the latter position—but it does so in a testable way, inviting us to prove it wrong.

16.2.10 How LoF coexists with optimization

Importantly, LoF (if true) does not replace RL or homeostasis; it constrains them. The Queue System (QS) introduced in Chapter 5 would act as a kind of meta-controller that prunes the choice set to policies with a high probability of neutral closure given the current ledger $L(T)$ and horizon H_t . Within that admissible set of options, ordinary optimization still runs its course: your habits, preferences, skills, and immediate incentives continue to determine which admissible action you take. In this picture, three practical corollaries emerge:

Agency is preserved: The individual still chooses among actions, but QS biases which actions are available or easy. It’s a nudge on the menu, not a puppet master deciding the final choice.

No miracles required: The “counter-balancing” experiences (if LoF is real) come via ordinary channels—sleep and dreams, personal insight, forgiveness, timely help from others, uncanny good-luck coincidences—rather than anything supernatural. They just occur more readily when needed and less when not needed.

Population coupling: The admissible menus are co-regulated across people so that my compensation does not steal yours. Chapter 5 discussed “coupled menus” – the idea that if two people are in deficit, the world won’t pit them such that one’s rescue dooms the other. In a LoF-governed system, behind-the-scenes constraints coordinate options to preserve a path to neutrality for everyone (a non-zero-sum fairness).

16.2.11 Where we go next:

In Section 16.3, we turn to composite rival models that blend reinforcement learning, homeostasis, and other processes. By constructing these hybrids, we ask whether familiar mechanisms alone could mimic life’s fairness-like patterns or if an actual fairness law must be at work.

16.3 Composite Rivals (Hybrids)

Single-mechanism theories rarely explain human life in full. The strongest rival to the Law of Fairness is therefore likely a hybrid that blends multiple processes: reinforcement learning (RL), homeostatic control, predictive coding/free-energy principles (PC/FEP), opponent-process adaptation, and social-ecological factors. Let's call this family of theories H*. An H* model claims that what looks like fairness in life is really an emergent tendency from many interacting parts—no terminal neutrality principle is required. In this section we charitably construct the best possible hybrids, then show where even these smart composites still diverge from LoF's strict path-wise closure prediction.

16.3.1 The *H* architecture at its best

A credible H* model includes at least five layers (each corresponding to a well-supported mechanism in neuroscience and behavioral science):

RL layer: Learns policies that maximize long-run returns under uncertainty, handling the exploration-exploitation tradeoff and possibly using risk-sensitive objectives (to account for risk aversion or risk-seeking).

Homeostatic layer: Regulates physiological and psychological set-points (sleep drive, nutrition/hydration, arousal level), minimizing deviation costs and preserving viability. This could include an allostatic component for predictive regulation.

Predictive coding layer: Compresses and explains sensory inputs by minimizing prediction error (or variational free-energy). In many PC/FEP models, affect (valence) corresponds to the confidence-weighted prediction error—surprising or high-uncertainty states feel negative. This layer would modulate behavior to reduce surprise.

Opponent-process and adaptation layer: Implements rebound and habituation phenomena. Pleasures “wear off” (diminishing returns) and pains “settle” over time. This biases the system toward homeostatic baselines after perturbations (the classic hedonic treadmill effect).

Social-ecological layer: Constrains and shapes options via external factors—scarcity of resources, cultural norms, institutions, and network effects (like reputation and support systems). This layer introduces phenomena like cooperation, punishment, caregiving dynamics, and other influences from the agent's social environment.

Promise of H*: With enough sophistication, these layers can explain repair (why people recover), resilience (why they often bounce back), and even some end-of-life shifts (since families and hospitals intensify care near death, etc.). H* can generate the appearance of balancing: after negative shocks, adaptation plus social support and re-optimization

can often move a person back toward typical ranges of mood. In short, H^* can claim that life tends to equilibrium because of all these buffering mechanisms.

Limits of H^ :* None of these layers, however, guarantees that the lifetime integral of felt valence for a single person closes at zero by the end. Each mechanism is either ensemble-wise or tendency-based. There is no explicit requirement that every individual trajectory neutralizes. Extreme, uncompensated paths can still occur in principle (and in practice).

16.3.2 How H tries to imitate LoF

To be a serious alternative, H^* must explain LoF's headline signatures (see 16.2.7) without invoking any terminal-neutrality constraint: horizon-dependent menu tilt, variance compression near death, and dream counterweights after hard days.

16.3.3 Where the best hybrids still diverge

Let's consider where even a top-tier H^* (with all layers included) will likely fall short of LoF, and how we could detect that:

(A) *Path-wise closure vs. expectation management:* All H^* mechanisms optimize expected outcomes, stabilize averages, and reduce errors. None of them forces the realized cumulative integral $\int_0^T F(t) dt$ for one life to be zero. Catastrophic sequences can still dominate an individual's experience before death. LoF is explicitly path-wise (each life individually closes neutral, barring measurement error); H^* is expectation-wise (on average things look balanced, but outliers can occur). *Empirical discriminator:* Track individuals with highly adverse trajectories (e.g. caregivers who suffer multiple compounded losses). Control for all known interventions (medical, social). If we observe an extra opening of reparative paths and unusual ease of finding closure precisely when horizons shrink—and these factors predict a neutral final outcome—then any H^* model would be forced to effectively include a closure constraint to account for it. In other words, if every high-variance life somehow finds a balancing turn at the end, H^* is missing something fundamental (and would have to smuggle in LoF to fix it).

(B) *Horizon coupling is selective, not generic:* H^* expects a generic effect like “people get more conservative when time is short.” LoF predicts a targeted effect: as time shrinks, specifically high- Φ (high-compensability) actions get facilitated and low- Φ actions become unavailable, even if their immediate payoffs are identical. It's not just caution—it's a precise filtering based on whether an action helps close the ledger. *Empirical discriminator:* Use a choice task where options are matched on immediate value, risk, etc., but differ in compensatory potential (say, one option offers emotional reconciliation, another just offers fun). Manipulate perceived horizon (e.g. priming

mortality or future scarcity of time). If LoF is correct, you'll find a $\Phi \times H$ interaction: the compensatory option's appeal jumps as horizon shrinks, beyond what any standard discounting or risk model predicts. This can be tested via both behavior and neural correlates (vmPFC, ACC, rIFG signals). If H^* lacks a fairness mechanism, it typically cannot produce that selective pattern—unless it cheats by adding a special term.

(C) *Population coupling without zero-sum loss:* As noted, H^* by default would predict that in zero-scarcity situations, those who are already worse off will get hit the hardest (their menus shrink first) because nothing ensures fairness. LoF predicts something more interesting: countervailing openings appear for those with the most negative ledgers, as if the system coordinates to preserve everyone's chance of recovery. *Empirical discriminator:* In real-world crises (economic recessions, disasters, competitive scenarios), look at whether people who have been suffering the most start to receive unexpected opportunities for relief (information, aid, reconciliation gestures) that disproportionately improve their situation relative to equally needy peers (after controlling for any formal aid policies). If data showed that, for example, in a housing crunch those who had been struggling the longest have a slightly higher rate of sudden windfalls or community support than would be expected, that would align with LoF's coupled-menus idea. If instead the worst-off consistently remain the worst-off or fall even further behind, LoF loses credibility on this front.

(D) *Dreams – inversion strength and conditionality:* H^* can certainly incorporate known dream functions like threat simulation and memory pruning. But LoF requires a very specific conditional effect: the worse yesterday was (affectively), the stronger the opposite-valence content in dreams, and the larger the next-day mood rebound—holding sleep quality constant. H^* as-is might explain some relief from dreaming, but it wouldn't necessarily produce a graded, deficit-dependent inversion.

Empirical discriminator: Do an intensive study with daily mood measurement and REM dream content coding (raters blind to the person's day). If the data show that the degree of positive (or comforting) material in dreams scales with how negative the prior day was, and that this in turn predicts the magnitude of next-day mood improvement—even after controlling for total sleep, REM duration, etc.—then LoF is capturing something beyond standard models. If those effects wash out with controls, then ordinary explanations suffice.

16.3.4 Building the strongest composite (H -Plus)

To give H^* every benefit of the doubt, imagine H^* -Plus, the most powerful non-LoF hybrid we can conceive. H^* -Plus would include:

Risk-sensitive RL (with meta-learning to handle non-stationary environments and adversarial agents).

Hierarchical homeostasis regulating not just physiology but also psychological variables (covering pain, sleep, inflammation, and social satiation levels).

Advanced PC/FEP with dynamic precision control (so affect tracks prediction errors, but the system can re-tune its priors under chronic stress to adapt).

Opponent-process calibration built-in, ensuring rebound and habituation happen on tuned timescales for maximum stability.

Network-aware social policy layer that actively allocates support to the worst-off (imagine a learned mechanism akin to a societal triage, making sure help tends to go where it's most needed, possibly via reputation or altruism signals).

Dream modulation module for threat extinction and autobiographical memory integration (basically a very sophisticated account of dreaming that can include some inversion of affect).

Even with all of this, what H*-Plus still cannot promise is almost-sure terminal neutrality for each stream regardless of unlucky draws. It can improve the odds of a life evening out (perhaps dramatically so, compared to a simpler world), but it cannot guarantee closure without effectively hard-coding something equivalent to the Queue System. For example, if multiple catastrophes strike during someone's final chapter of life, H*-Plus has no inherent law that says "and still, somehow, they will find balance before the end." Only a LoF-like constraint ensures that.

16.3.5 Adversarial fit and overfitting risks

A sophisticated H* (or H*-Plus) can always be tuned to match observed data post hoc. If LoF proposes a signature, a determined modeler could bolt on mechanisms to mimic it. Two concerns arise here:

Goodhart at theory level: Once a target phenomenon (say, end-of-life variance compression) is widely known, modelers might introduce ad hoc penalties or couplings in their H* models to produce that outcome, even if those additions aren't independently justified by theory. In essence, the model could "cheat" by implicitly adding LoF-like terms just to win the contest, rather than because nature truly operates that way.

Generalization: A hybrid model that is tuned to match one context (e.g. hospice end-of-life data) must also reproduce the other signatures in other contexts (horizon effects in mid-life decisions, dream counterweights after everyday stress, coupled-menu effects

in scarcity) without retraining. If you overfit to one, you might fail on the others. A true LoF mechanism, by contrast, should manifest across all these contexts with one unified principle.

To guard against these pitfalls, a fair test should involve pre-registered adversarial collaboration (as we outline in 16.4). Some key steps: (i) train each model on a subset of contexts and evaluate on others (to test generalization), (ii) have independent teams swap data sets (to avoid analytic bias), (iii) include “hostile splits” where only the hardest conditions (e.g. very short horizons with tempting options) are held out for final testing. The idea is to prevent any side from overfitting just one scenario.

16.3.6 Minimal add-ons that turn H into LoF (stealthily)

It’s worth noting there are only a couple of ways one might “smuggle” a fairness constraint into an H^* model without calling it LoF:

Horizon-dependent utility shaping: Add an extra term in the agent’s utility function that grows large (penalizing unresolved negative ledger) as $H \rightarrow 0$. In effect, this is a horizon-indexed cost for dying with a big negative balance. This is mathematically equivalent to giving the agent a shadow price λ_t on suffering that increases near the end. That λ_t term is basically what LoF posits (see Chapter 6 on “shadow price”).

State-dependent action masking: Program the model such that certain action sequences get forbidden if they would push the cumulative affective total below some threshold near the end. In other words, as the ledger grows more negative and time is short, the model quietly removes those options. That is precisely the Queue System’s pruning rule.

If a composite model requires either of these mechanisms to match the data, then it has become LoF in all but name. It’s no longer a pure rival; it’s a stealth implementation of the very thing it was arguing against.

16.3.7 Practical tests: LoF vs. the best hybrids

Building on the earlier discussion, we can propose a few practical experiments or analyses to decisively separate LoF from even a very strong H^* :

Selective tilt test: Design a decision-making experiment where immediate rewards are equal across options, but one option has high long-term compensatory value (e.g. mends social bonds) and another does not. Manipulate the “horizon” (e.g. tell participants they have only one more day or many years). Prediction: Under LoF/QS, when the person’s ledger is negative and the horizon is short, the reparative option will gain extra value and be chosen much more often than in long-horizon or neutral-ledger

conditions. We'd measure a significant $\Phi \times H$ interaction effect in choice probabilities (and possibly neural decision signals). Controls include ensuring any differences aren't due to basic risk aversion, discounting, pain levels, etc. A consistent finding of this selective tilt (across participants and contexts) would favor LoF. Lack of it would favor H^* .

Coupled-menu test: In a competitive or resource-scarce environment (a simulated economy or multiplayer game), identify individuals with similar needs but different cumulative "suffering" histories. Introduce a new limited opportunity for relief or improvement (like a sudden job opening, or a helping hand from a stranger). Prediction: LoF implies that those with more negative ledgers are more likely to encounter or seize these spontaneous repair opportunities (even beyond what their skill or effort alone would yield). We would need to equalize policy inputs (so everyone is trying as hard and is as skilled) and see if the worst-off still get some extra breaks. If data show that opportunities distribute purely according to external factors or even against the worst off, LoF loses credit. If there's a subtle bias for those in long-term deficit (that can't be explained by known aid or by them trying harder), it suggests an underlying coordination (pointing to QS-like coupling).

Dream conditionality test: As mentioned, sample many nights of dream reports and next-day mood, ideally in a population where we also manipulate "horizon" cues (e.g. some nights participants are reminded of their mortality or another horizon-shortening context before sleep). Prediction: Under LoF, the degree of affective inversion in dreams (how positive the dream is after a negative day) should correlate with how negative the prior day was and be stronger when a person's overall horizon context is limited (e.g. during an emotionally "critical" period). Also, those dreams should predict a quantifiable rebound in next-day Hedonic Composite Index (HCI) score beyond normal. If this holds, especially under rigorous controls (sleep architecture, blind content coding), it bolsters LoF. If dream effects reduce to generic processing (e.g. everyone just has proportionate dreams to daily events, with no special inversion for big deficits), then H^* can explain it.

Results that consistently favor LoF (across multiple independent labs and populations) would strengthen the case for a true fairness law. Conversely, if the best H^* models match all observations without needing any closure tricks—and negative results accumulate on those key tests—then the composite view wins.

16.3.8 Summary for readers (and modelers)

We have given the hybrid models their strongest form. These H^* composites incorporate the best of modern neuroscience and behavioral science; they do explain a lot about competence, stability, adaptation, and why life often self-corrects. But they stop short of

what fairness—in any morally serious sense—demands: closure in each life, not just on average in the aggregate. If a composite H^* can truly reproduce all of LoF’s predicted signatures without importing a closure constraint, then it deserves to win as the better theory. If it cannot, then the evidence will indicate that a law-like admissibility constraint (the Queue System idea) belongs in our fundamental description of how conscious lives unfold.

16.3.9 Where we go next:

Section 16.4 will formalize the head-to-head tests between LoF and its rivals. We outline an adversarial collaboration framework—pre-registered experiments and model comparisons—that ensures each side gets a fair trial and that the Law of Fairness faces genuine opportunities to be proven wrong.

16.4 Research Notes: Model Comparison and Adversarial Fits

This section outlines a playbook for fair, head-to-head testing of the Law of Fairness (LoF) against the strongest composite rivals (H^* from 16.3). The guiding principles below aim to prevent “victory by clever modeling tricks” and instead reward pre-registered, out of sample, multi-site performance on the signatures that matter most.

16.4.1 Competing model families

We define two broad model families to be compared:

LoF–QS (constraint model): This is LoF instantiated via the Queue System mechanism. Core terms include the running ledger L_t , the horizon H_t , and a compensability score $\Phi(u; L_t, H_t)$ for an option u given current ledger and horizon. LoF–QS predicts a selective horizon \times compensability interaction in choices and corresponding admissible-set pruning (i.e. some options disappear or become non-viable when they would jeopardize terminal neutrality). In practice, a LoF–QS model might look like a decision policy with an extra term or filter enforcing near-future neutrality.

H^* (composite rivals): This family encompasses the best hybrids combining RL, homeostasis, PC/FEP, opponent-process dynamics, and social-ecological factors (see 16.3). These models can capture adaptation and resilience via known mechanisms without invoking any explicit terminal neutrality principle.

Each family must be fully specified before seeing the test data. That means fix all parameters, priors, learning rules, and preprocessing steps in advance (ideally in a Registered Report or similar protocol). No tweaking after peeking!

16.4.2 Outcomes and signatures

We will evaluate models on three tiers of phenomena (as introduced earlier):

Micro-choice signatures (within-individual effects): Primary effect: a significant $(\Phi \times H)$ interaction in decision outcomes and neural decision signals. For example, in an experiment, when horizon is shortened, do people and their brains disproportionately shift toward high- Φ (compensatory) options? We will control for utility, conflict, arousal, discounting, risk, analgesia, etc., to isolate this effect. LoF–QS expects this interaction; many H^* variants do not, unless they include a similar term.

Mesoscale dynamics (coupled menus under scarcity, between individuals): Primary effect: preferential opening of reparative paths for “high-debt” streams beyond what any policy inputs or initial advantages would predict. This is the coupled- menu tilt discussed: do those with larger negative ledgers somehow end up with relatively more

opportunities for relief? We measure things like the probability of a positive event (reconciliation, surprise aid) as a function of one's cumulative past suffering, controlling for obvious factors.

Macro trajectories (within-person trajectories near end-of-life): Primary effect: within-person variance compression and drift toward neutrality that cannot be explained by standard clinical or reporting factors. We look at whether individuals' HCl (hedonic index) variance decreases and mean moves closer to zero in their final phase, compared to earlier in life and compared to matched controls. Key is controlling for sedation, analgesia, ritual, etc. If an extra narrowing remains, that's LoF's signature.

Dream counterweights (as an auxiliary battery of tests): We will also examine the valence inversion phenomenon in dreams (Chapter 10) – specifically, whether negative-day deficits predict opposite-valence dream content and improved next-day mood, holding sleep architecture constant. This is a supportive test: LoF says “yes, you'll see that”; H* might allow some effect but not as a tight function of the deficit.

All these outcomes will be quantified in a way both models can attempt to predict (e.g. via likelihoods or simulations).

16.4.3 Design: preregistration and adversarial collaboration

To ensure rigor and objectivity:

Preregistration (Registered Reports): We publish detailed protocols in advance. This includes the exact model code for each side, priors, which features and nuisance regressors we'll include, exclusion criteria, and what counts as success. No fishing expeditions—everything is locked in.

Adversarial teams: We form two teams (or more) – e.g. one championing LoF-QS, one championing H*. They co-author the experimental protocol, agreeing on fairness. Each team freezes their model parameters and analysis code ahead of time. We also implement a cross-analysis step: each team will analyze the other's held-out data to check that results aren't an artifact of analysis choices. This forces methods to be robust.

Multi-site, multi-population: To avoid any “lab-specific” or culture-specific outcome, we run studies in at least three independent sites with diverse populations (different demographics, maybe different countries). Different MRI scanners or survey platforms, different experimenters. We include site indicators in analyses and use partial pooling (hierarchical modeling) so we can detect any idiosyncratic site effects. The goal is to ensure findings aren't a fluke of one setup. Also, before doing cross-cultural comparisons, we will verify at least configural and metric invariance of the key measures

across groups (and scalar invariance if sample sizes allow) – in other words, make sure HCI or self-report scales mean the same thing across sites before pooling data or comparing means. (If invariance fails at scalar level, we restrict to within-site patterns or use latent-variable methods that account for intercept differences.)

16.4.4 Data partitions and leakage control

To fairly assess generalization:

Immutable split: We designate certain sites or subsets as training/validation (say Sites A + B) and another as the final test (Site C). The models can be trained and tuned on A+B, then locked. All confirmatory tests are then run on C, which neither team has touched until that point. This ensures what we report isn't just overfitting to the idiosyncrasies of one sample.

Hostile split: Additionally, within our data, we reserve the hardest scenarios exclusively for the final test. For example, maybe in Sites A+B we include moderate horizon lengths, but in Site C we include some very short horizon with high temptation scenarios that really stress the models. Neither model gets to practice on those extreme cases until the final evaluation. This is to see if LoF (with its extra constraint) shines exactly in those tough cases, or if H* can extrapolate.

Leakage guards: We take data integrity seriously. Outcome coders (e.g. people rating dream content, or identifying reconciliation events) are kept blind to conditions and hypotheses. We use containerized pipelines and cryptographic hashes for all raw and processed data to ensure no accidental (or intentional) tampering. Essentially, we want to rule out p-hacking or subtle biases creeping in.

16.4.5 Model fitting and comparison

We outline how each model will be fit and compared statistically: For behavioral choices, we'll likely use hierarchical logistic regression for binary decisions or softmax (multinomial logit) models for multi-option choice sets, with subject-level random effects to capture individual differences. For example, LoF-QS might predict choice probability using utility plus a $\Phi \times H$ interaction term (and, where specified, admissible-set pruning), whereas an H* model might use utility and other state variables but omit any explicit horizon-dependent $\Phi \times H$ interaction. We'll examine how well each fits choices.

For neural data (fMRI, EEG, etc.), we focus on key regions of interest (ROIs) implicated by LoF: rIFG, ACC, vmPFC, insula (for example). We use GLMs with parametric modulators (e.g., modeled value terms, Φ , horizon, or their interaction), then perform hierarchical

pooling of those effects across subjects. Specifically, we test whether BOLD responses in these ROIs track the predicted $\Phi \times H$ interaction in modeled value estimates under short horizons in LoF versus H^* .

We choose priors that are weakly informative (to stabilize estimation without biasing toward a particular effect). We will also conduct sensitivity analyses to ensure results are not artifacts of particular prior choices.

For metrics of fit: We'll use modern information criteria like PSIS-LOO (Pareto-smoothed leave-one-out) cross-validation and WAIC (Watanabe–Akaike Information Criterion) to assess pointwise predictive accuracy. These essentially tell us which model predicts new data better, while penalizing complexity. We'll report things like stacking weights (how to weight models in an ensemble for best prediction). For simpler subsets we might also compute Bayes factors via marginal likelihood (with bridge sampling) if feasible, though for large complex models we'll rely on LOO/WAIC.

We'll also perform posterior predictive checks (PPCs): simulate data from the fitted models and compare to the actual data distributions. For instance, does the LoF model, when simulating many “worlds,” produce the kind of $\Phi \times H$ slope we saw empirically? Does the H^* model produce it or not? We will specifically check distributions of: (i) the $\Phi \times H$ interaction effect sizes, (ii) “menu shrinkage” indices (some measure of how much someone's choice set contracts near end-of-life), (iii) end-of-life variance in affect, (iv) dream inversion metrics. The models should reproduce those if they're correct.

We'll evaluate calibration and discrimination too. Calibration: for example, if a model says 70% chance someone will choose a relief option, does that happen ~70% of the time? Discrimination: use metrics like Brier scores or ROC–AUC for binary outcomes (e.g. did a person initiate a reconciliation or not by end-of-life). A model that both predicts and is well-calibrated gains credibility. (As a note, any count-based outcomes like “number of options considered” will be modeled with appropriate distributions: we'll use Poisson regression for counts and explicitly check for overdispersion. If dispersion > 1.2 , we'll switch to a Negative Binomial model to account for extra-Poisson variance. This ensures, for example, that if we measure something like the number of spontaneous repair opportunities in a week (a count), our model inferences aren't mis-specified.)

16.4.6 Ablation and “knock-out” logic

To probe why a model succeeds or fails, we plan structured ablations:

For the LoF–QS model, we will systematically remove its key components and see what happens. For example: remove the Φ term (no compensability factor), or remove horizon H_t coupling (no special end-of-life sensitivity), or remove the admissible-set masking

entirely. If LoF is truly essential, we expect a stepwise collapse of its predictive power and its ability to produce the signatures when we strip those parts. (This is effectively testing if QS's elements are really what drive fit.)

For the H* model, we do the opposite: strip away each layer (no opponent-process layer, or no explicit homeostatic cost function, etc.). If none of those removals significantly reduces a $\Phi \times H$ effect (because originally H* might not have one at all), that's telling. If H* only shows a fairness-like effect when, say, we had added a horizon penalty, then that ablation will reveal it (removing that penalty will lose the effect).

Interpretation rule: if an H* model only fits the data well after adding something like a horizon-dependent penalty or an action mask, then effectively it has imported LoF's mechanism. We would count that against H* as an independent explanation (because it needed LoF's crutch to work).

16.4.7 Specification-curve and multiverse analyses

We will not hide alternative analyses in a file drawer. Instead, we enumerate all reasonable analytic choices: e.g. how to preprocess fMRI (motion threshold? which hemodynamic response basis?), how to binarize dreams (by theme? by scorer?), what discount- rate prior to assume, how to code “reconciliation” in hospice notes, etc. Then we perform a specification curve (a.k.a. multiverse analysis). This means we run the analysis under all combinations of those reasonable choices and display the distribution of outcomes.

This will show which findings are robust (they appear consistently across choices) vs. which are fragile (they only appear under certain specifications). Only effects that survive the multiverse with at least qualitatively consistent size/direction will be considered strong. If an effect disappears with slight tweaks, we'll be transparent about that (and likely not lean on it as evidence).

16.4.8 Out-of-sample stress tests

Beyond the initial train/test split, we subject models to additional “transfer” tests: Context transfer: We might train models on lab task data (e.g. short-term decision experiments) and then test their predictions on longitudinal hospice telemetry data (end-of-life monitoring), and vice-versa. Does a model that learned on one scale of observation predict the other? For LoF to be convincing, a model incorporating it should handle both contexts reasonably, whereas a rival might stumble in the drastically different setting unless it truly captured the underlying principle.

Population transfer: Test models trained on one demographic or clinical group on a very different group (young healthy adults → older chronic pain patients, etc.). LoF's mechanisms are meant to be fundamental, so they should show up across populations if true (once measurement invariance is confirmed).

Temporal transfer: Train on early-life adversity cohorts, test on their midlife outcomes, or train on midlife, test on late-life. Essentially, can the model generalize across time scales?

We will define success criteria a priori. For example: a model “passes” if its LOO predictive accuracy on the transfer is within some acceptable range (or if a LoF model’s advantage remains > 10 LOOIC points on new data, etc.). We might also set criteria like “the effect size under LoF remains at least X and credible interval excludes 0 in the new sample,” etc. The idea is not to retroactively cherry-pick what counts as success—those thresholds will be agreed upon in the preregistration.

16.4.9 Negative controls and falsification traps

We build in negative control experiments to catch any spurious “fairness” effects:

Pseudo-horizon controls: We can introduce fake “time is running out” cues that do not actually change a person’s true horizon. For example, a game where we tell them “this is the final round” but actually it’s not, or use background cues that suggest finality but then continue. LoF predicts no special Φ tilt if the horizon isn’t truly shorter (i.e. the brain shouldn’t tilt choices if it later finds out time wasn’t actually short). A generic H^* might still show behavior change because the person believed time was short (risk aversion kicks in generically). Observing a tilt in the fake scenario would indicate the effect was more superficial (just psychology of perceived time) rather than an actual QS mechanism, which would be a strike against LoF’s uniqueness.

Permutation dreams: Shuffle the pairing of dream reports and days for analysis (a classic permutation test). If we still see an “inversion” correlation (yesterday’s negativity with tonight’s positive dream) after shuffling days, then it’s an artifact or bias in coding. A real counterweight effect should vanish or weaken under such permutation.

Medication confounds: In end-of-life data, include time-varying covariates for analgesic and sedative medications (with lags, since drugs take time to act). LoF’s predicted variance compression or neutrality should persist after accounting for pharmacological effects; if it doesn’t, then perhaps drugs, not LoF, explain the effect.

We predefine Fail patterns for LoF: for example, (i) if after proper controls we see no $\Phi \times H$ interaction anywhere, (ii) if we see no shrinkage of options near terminal times in any

dataset, (iii) if H^* models can fit just as well without any closure proxies. These are essentially conditions under which we'd consider LoF to have lost the contest.

Fail pattern: No detectable $\Phi \times H$ tilt. After accounting for known factors, if short horizons do not lead to any selective bias toward high- Φ (high-compensatory) options in behavior or brain, then QS isn't manifesting. This would be a major failure for LoF's micro-level prediction.

Fail pattern: No menu narrowing at end-of-life. If individuals nearing death (in well-supported conditions) show no drop in option variety or no shift toward relief/closure actions—i.e. they continue pursuing all types of actions as before, and some die with large unaddressed “ledger” disparities—then LoF's core claim is unsupported.

Fail pattern: Rival fit is equal without closure. If the composite H^* models achieve equal or better predictive accuracy on all key measures without adding any horizon-dependent penalties or action masking, then LoF has not proven its necessity. The rivals would explain the data with ordinary terms, undermining the case for a new law.

(If any of the above patterns emerge in two or more independent tests, we would consider LoF essentially falsified under those conditions.)

16.4.10 Reporting thresholds and decision table

We will establish clear criteria for declaring one model the “winner.” Primary endpoints include:

Micro-level: A significant, directionally consistent $\Phi \times H$ effect on choices and in neural ROI signals, replicated across sites with proper multiple- comparison correction. (E.g. if two of three labs see a robust effect and one is borderline, that might still count as consistency if pooled analysis confirms it.)

Macro-level: A within-person affect variance compression near end-of-life that exceeds that seen in matched controls (e.g. patients of similar condition who are not near end) and beyond what clinical covariates explain. Also, a final mean affect that lies within the pre-registered neutrality band ($\pm 0.15 z$) with a slope near zero.

Dreams: A demonstrated link “prior-day deficit → opposite-valence dream content → next-day mood rebound,” independent of sleep quality.

Then, model victory conditions:

We'll say LoF “wins” if it clearly dominates in predictive performance (lower WAIC/LOO scores by a meaningful margin) and it accounts for those primary endpoints in ≥ 2 out of

3 sites (assuming we have 3) and shows it can transfer to at least one distinct context. Essentially, LoF has to win on both explanation and prediction in the majority of cases.

H* “wins” if it matches or exceeds LoF’s predictive fit without needing any closure proxies and it passes all the negative control checks (meaning any $\Phi \times H$ type effects we saw could be explained by its mechanisms or were artifacts).

All code, all preregistrations, and de-identified data will be made public (in registries or repositories) once the initial results are verified. The emphasis is on transparency: anyone should be able to reproduce the analyses and test alternate models on the same data.

16.4.11 Adversarial fits in practice – a minimal protocol

Putting it together, a likely real-world sequence:

Stage 1 (pre-data): Teams publish a joint paper detailing background, hypotheses, and frozen model code + analysis plan. This gets peer-reviewed (perhaps conditionally accepted) before data collection.

Pilot phase: We run small pilots to calibrate task difficulty, ensure measures have enough dynamic range, and refine any logistical issues (e.g. are people understanding the instructions? Is HCl variance measurable on the timescale?). Pilot data might also help choose priors but is not used for hypothesis testing.

Concurrent data collection: All sites collect data in parallel according to the protocol. We include midstream audits to ensure nobody is deviating from the preregistered plan (e.g. all labs must follow the same scripts, etc.).

Locked analysis: Once data is locked, each team runs their analysis on their own data and also on the other’s (if we withheld some data for cross-check). Analyses are ideally automated in a pipeline with no human tweaking.

Joint interpretation: Both teams then meet (or exchange results) and populate a decision table (as defined above) that tells us which conditions were met. We then draft a joint report interpreting the outcomes strictly through the lens of the pre-set criteria (to avoid post-hoc re-framing).

Replication tranche: If LoF emerged looking good, one might then invest in a replication with roles reversed (e.g. swap which site was held-out vs. training) to verify generalization. If H* looked better, perhaps try a new context to poke once more at LoF’s weak spots. In any case, we’d likely plan a follow-up where, say, what was test becomes train and vice versa, to double-check any site or cohort effects.

16.4.12 Why this rigorous approach matters

LoF is a claim of a law, whereas H* is a claim of a strong tendency. The only way to distinguish a true law from a convincing tendency is rigorous, adversarial, predict future experiments and model comparisons. If LoF can survive this kind of gauntlet—producing specific, quantitative predictions that hybrids cannot match—then it will have earned the right to be treated not as metaphor or wishful thinking but as a genuine governing constraint on conscious minds. If instead the H* models prevail without having to sneak in any closure mechanisms, then LoF will have learned exactly where nature stops short of true lawhood. Either outcome pushes science forward: we either validate a new fundamental law or we precisely delineate the limits of existing theories, absorbing the lessons into improved models of mind and behavior.

16.4.13 Where we go next:

Having tested the Law of Fairness against its strongest challengers, we now turn to the broader canvas of life. In Act VIII, we examine how evolution and social dynamics play out under this constraint. Chapter 17 asks how natural selection might “meet” a fairness law, outlining the non-negotiable “no-go zones” in phenotype space and the everyday strategies organisms use to skirt them without invoking any cosmic purpose.

16.5 If Rivals Win, What LoF Learns

What if, after all this, the best composite rivals (RL + Homeostasis + Predictive Coding + Opponent Process + etc. hybrids) consistently out-predict the Law of Fairness in our strong tests? This section is our contingency plan—an honest account of how the LoF research program would adapt if nature says “no” to a strict fairness law.

16.5.1 Theoretical downgrades: from law to regularizer

If the rivals clearly win without having to import any closure proxies (meaning no hidden horizon terms, no QS-like action masking in their models), then we would downgrade LoF from a proposed law to a more modest hypothesis:

LoF → “Fairness Regularization” Hypothesis (FRH): We would reinterpret the “Queue System” not as a fundamental governing constraint of the universe, but as an emergent, heuristic manifold that some agents approximate under certain conditions. In plainer terms, the $\Phi \times H$ interaction (compensability by horizon) would be seen as something that often emerges via learned behavior and social scaffolding, rather than an inviolable law of nature. It might still be a useful concept, but not nomologically necessary.

Ledger $L(T)$ → affective load index: We’d treat the cumulative ledger not as a conserved quantity that must zero out, but simply as a useful summary metric (like allostatic load or stress burden) that tends to correlate with outcomes. It could still be valuable for predicting who needs help, for instance, but we wouldn’t claim it inevitably goes to zero.

Admissible set $\mathcal{A}(t)$ → feasibility cone: Rather than imagining a mystical narrowing of choices by a fairness law, we’d attribute menu changes to the intersection of resource limits, uncertainty, and social constraints. We might still study how people’s feasible options change near death, but frame it as “given what they can do (energy, help, knowledge), here’s what tends to happen” instead of implying a cosmic referee.

End-of-life neutrality → variance artifacts: What looked like people “settling at neutral” at end-of-life we’d reinterpret as likely due to known factors (pain management, sedation, ritual, expectancy effects, caregiver behavior) unless we found any residual beyond those. In other words, unless some of that effect survived all controls, we’d chalk it up to a combination of those mundane factors and drop the idea of a hidden force guiding it.

In summary, LoF would be reframed not as a law, but as a potential regularizer or design principle that evolution and society might approximate in some ways, but not an ironclad rule.

16.5.2 What still stands as a contribution

Even if the law itself fails, the project is not for naught. Several things we've built would still be valuable:

- Measurement framework: We introduced a Hedonic Composite Index (HCl) and Hedonic Composite Units (HCU), and hammered on measurement invariance (configural → metric → scalar), rigorous preregistration, blinding, etc. These are field-advancing standards for quantifying subjective experience without self-deception. Even absent a LoF, having a solid “common currency” for affect and a robust way to measure it across people is a big win for affective science.
- Task design and analysis norms: We developed things like horizon-manipulated choice tasks, admissible-set probes, QS-residual modeling after accounting for other factors, multi-site Bayesian workflows (with LOO, PPCs, stacking weights), and spec-curve reporting. These raise the bar for rigor in this field. Future studies on well-being or decision-making can borrow these methods to be more credible, whether or not they test LoF per se.
- Dream methodology: We created a clear protocol for paired day–dream–next-day analysis with sleep architecture controls. Even if we conclude that “counterweight dreams” aren’t a special fairness mechanism, this methodology can still be used to explore how dreams relate to emotional processing (be it memory consolidation, threat simulation, or otherwise). We’ll better understand dreaming in any case.

So regardless of LoF’s fate, these tools and standards remain valuable contributions.

16.5.3 Negative results that would force the retreat

We commit up front that we will update our claims (and tone them down) if we robustly observe certain outcomes. Specifically, if we repeatedly find:

- Null $\Phi \times H$ effects in behavior and in key brain regions (vmPFC, ACC, rIFG, insula) even with large samples and proper controls. If compensability-by-horizon just isn’t showing up beyond noise, that undercuts QS.
- No admissible-set shrinkage near objectively short horizons when obvious reparative options exist. For example, if in an experiment people near “the end” (of the scenario) still pursue trivial or even harmful options with equal ease as earlier, then there’s no hidden pruning helping them out.

- No variance compression in end-of-life affect after adjusting for medication, sedation, self-report bias, etc. If people's mood variance and drift at end-of-life remains as wide as ever (just maybe lower on average due to morphine or such), then LoF's strongest empirical prediction fails.
- Rivals win generalization tests: If models trained on one context (lab tasks) generalize better to another context (hospice data) without needing LoF terms, or vice-versa, and LoF-augmented models do not improve predictive power, then LoF isn't earning its keep as a principle.

If we see those findings robustly (multiple sites or very decisive single-site evidence with high power), we will cease to claim lawhood. In practice, that means our publications and communications would switch to saying “we did not find evidence of a strict Law of Fairness; instead we found evidence for XYZ processes.”

16.5.4 How LoF integrates rivals' lessons

If the hybrids prevail, we won't just walk away empty-handed. We will actively integrate what we've learned into the existing frameworks:

- *In a predictive coding (FEP) context:* perhaps we'd recast our Φ (compensability factor) as something like a compound precision term on predicted affective error under uncertainty. The “pruning” of menus might be reframed as the system adjusting precision weights and model uncertainty – effectively, the brain selecting policies that minimize long-horizon prediction error rather than enforcing fairness per se. This ties QS to established PC principles.
- *In RL + homeostasis terms:* we might reinterpret “ledger dynamics” as a form of allostatic load management. That is, what we thought of as balancing the ledger could be seen as the organism taking actions that reduce long-term deviations from viability set-points (which sometimes masquerades as fairness). “Repair” options might just be those that reduce anticipated future allostatic cost. We'd effectively absorb LoF phenomena into an expanded homeostatic/allostactic model.
- *For opponent processes:* the rebound effects we studied can be viewed as the nervous system's natural push toward equilibrium after extremes, without invoking fairness. Our contribution there might simply be that we helped quantify these rebound dynamics more precisely (e.g. by using composite metrics rather than single scales). The concept of “counterweights” could remain as a metaphor, but we'd attribute them to known neural equilibrium mechanisms, not a new law.

In short, if LoF fails, we will retrofit its key ideas into the language of the successful rivals, ensuring nothing valuable is lost—only reinterpreted.

16.5.5 What practical guidance survives anyway

Even without a law, much of the practical advice derived from LoF would remain unchanged because it was beneficial regardless:

- Clinical translation: The “menu widening” strategies we advocated (e.g. aggressively controlling pain, stabilizing sleep, ensuring social support, facilitating closure conversations) are good palliative practices in their own right. They help patients feel better and make use of their remaining time more meaningfully, whether or not a neutrality law is guaranteed. We would still urge caregivers to follow those principles—LoF or not, they improve lives.
- Policy and design: We emphasized that one person’s options can constrain another’s (the coupled menu idea). Even if fairness isn’t cosmically enforced, it’s ethically and practically wise to design policies that recognize interdependence. For example, resource allocation and conflict mediation should consider how to avoid trapping a subset of people in no-win scenarios. Fairness as a guiding principle (if not a law) can still inspire better institutions (e.g. restorative justice, social safety nets that ensure people have paths to recover).
- AI and simulation engineering: The Queue System abstraction—an admissible-set filter that blocks policies likely to lead to large aggregate suffering—remains a useful engineering pattern for AI safety. Even if nature doesn’t guarantee fairness, we might want our AI or virtual worlds to have something like that for ethical and stability reasons. So implementing a QS-like mechanism in software (to avoid scenarios that produce extreme suffering unless compensations are in place) is still a positive contribution from this work.

In summary, the ethos and many applications of LoF survive, even if its status drops from “universal law” to “interesting heuristic”.

16.5.6 A decision table for claims going forward

We can summarize the possible outcomes and how we’d adjust our stance:

- Strong $\Phi \times H$, clear menu shrinkage, terminal compression observed (with out-of-sample predictive wins for LoF) Status of LoF: Supported Recommended stance: Maintain that LoF is a plausible governing law of experience. Continue law-level claims (with appropriate caution).

- Mixed results (e.g. micro-level effects seen, but macro neutrality weak; or rivals match LoF on fit without closure proxies in some aspects) Status of LoF: Weakly supported
- Recommended stance: Downgrade to Fairness Regularization Hypothesis. Treat LoF as a useful organizing principle or emergent regularity rather than a strict law.
- Robust null findings across tiers, rivals dominate predictive tests. Status of LoF: Not supported
- Recommended stance: Retire the “lawhood” claim. Acknowledge no evidence for a universal law. Still emphasize measurement contributions and explore fairness phenomena as emergent features.

This table will guide how we phrase conclusions in publications. We will not oversell if we land in the weak or not-supported zones.

16.5.7 New hypotheses to pursue if LoF loses

Even a negative result opens new questions. Some ideas to explore next:

- Conditional fairness: Perhaps fairness neutralization is real, but conditional – appearing only in certain subpopulations or contexts. For example, extremely tight-knit communities or high-social-support contexts might exhibit quasi-neutralization (via culture and norms) even if the universe at large doesn’t enforce it. We could identify those conditions (ritualized reconciliation practices, etc.) and model fairness as an emergent socio-computational phenomenon rather than a basic law. Essentially, study fairness as something societies or systems achieve under certain conditions (a kind of contingent equilibrium).
- Network fairness: Maybe fairness is better framed at the network or group level – e.g. as a graph-theoretic regularizer on coupled agents. Instead of each individual guaranteed closure, the idea would be that the network of interactions tends toward balance in aggregate. We could formalize something like “no subset of the graph consistently parasitizes the rest without consequences”. This might explain why “my menu depends on your choices” under scarcity—fairness might be a property of equilibrium in a coupled system, not an individual law.
- Dream roles revised: If counter-balancing dreams don’t hold up as a fairness mechanism, we’d test alternative theories: e.g. do dreams primarily reflect prediction-error pruning (reducing over-precise predictions) or memory integration rather than hedonic balancing? We could do more invasive tests, like closed-loop sensory stimulation during REM to see if we can alter dream valence

and then measure next-day affect. This would help separate the “counterweight” hypothesis from a straightforward memory consolidation effect.

In short, a null on LoF doesn’t end the inquiry—it refocuses it on where hints of balance come from in a world without a law.

16.5.8 Ethical posture, unchanged

We want to emphasize: whether LoF stands or falls, our ethical guidelines remain exactly the same. We never intended LoF to justify harming anyone or doing anything reckless. So:

- No one should attempt any dangerous “falsification” stunts (e.g. trying to cause suffering to see if it balances—this was always off-limits and remains so).
- Absolutely no end-of-life interference or “terminal manipulations” under the guise of testing anything.
- Informed consent, dignity, and compassion for all research participants (and all conscious beings we study) remain paramount.
- Strict guardrails remain on telemetry and simulation work (e.g. if we simulate conscious-like agents, we treat even the possibility of their suffering with extreme caution).
- Scientific disappointment does not grant ethical license. If LoF fails, it might be emotionally disheartening, but it does not permit us to loosen any ethical standards in pursuit of something else.

In summary, none of the humane protections we’ve insisted on change one iota based on LoF’s truth or falsity. Those were grounded in basic ethics, not in LoF’s validation.

16.5.9 What would still count as a “partial win”

Even if we cannot declare a Law of Fairness, the effort could still achieve meaningful pieces of its vision:

- We might establish measurement invariance across diverse populations for subjective well-being and thereby produce a field-ready composite affect index (HCI). That alone is a big accomplishment—making “happiness” or “suffering” something we can measure and compare validly (Chapter 8’s ladder). That helps science even without a law.
- We might demonstrate horizon-sensitive control as a common structure in brains and behavior. For instance, even if not strong enough to enforce neutrality, we

might find that many brains implement a “safety brake” as time runs out (maybe to ensure at least some goals are met or to avoid regrets). Documenting that mechanism (in ACC/rIFG etc.) is valuable in itself.

- We could deliver admissible-set engineering techniques that prove useful in clinical settings or AI. For example, the idea of pruning away options that would lead to huge negative outcomes can be applied in therapy (nudging patients away from self-destructive choices) or AI safety (preventing algorithms from actions that could cause massive harm unless certain conditions are met). That engineering approach is helpful with or without a natural law.
- We will have mapped the conceptual terrain thoroughly: by seeing exactly what aspects of LoF rivals had to augment to fit data, we clarify what new concepts or interactions are actually needed for a full account of affect dynamics. Even if those concepts end up folded into existing theories, we’ve charted them.

Any of those outcomes would let us claim some success. LoF’s bold claim might not fully hold, but we’d have advanced the science of consciousness and well-being in tangible ways.

16.5.10 Closing note: why this section exists

It might seem odd to devote a full section to “what if we’re wrong.” But this is precisely how a theory earns trust: by either surviving its toughest tests or by gracefully transforming in light of them. We want to make it clear that we are not invested in LoF being true at all costs. We are invested in discovering the truth, whatever it may be. If the rivals win cleanly, we will say so plainly, keep using any tools that proved useful, and redirect our efforts to understanding how fairness-like outcomes emerge in brains, societies, or machines without a law. If LoF instead survives the gauntlet of tests we’ve laid out, that same adversarial rigor is what will make its victory convincing to the scientific community. Either way, the science moves forward.

16.5.11 Where we go next:

In Part VIII, we step outside the internal tests and ask how a fairness constraint would manifest in nature and design. Chapter 17 (“Natural Selection Meets a Law”) probes how evolution might cope with or reflect a built-in balance constraint across species and cultures, and Chapter 18 (“If Life Is a Game”) explores the idea from a simulation standpoint, treating life as a designed game-world to see if a fairness rule would make such a world stable. Having stress-tested LoF against its rivals, we now investigate its implications in evolutionary processes and engineered worlds.

Part VIII — Evolution and Simulated Worlds

Life did not have to become what it is; it could have been otherwise. Yet across billions of years, organisms converge on familiar motifs—nervous systems that track harm and benefit, homeostatic controls that keep bodies within narrow bounds, social structures that police freeloading and reconcile conflict. In the framework of this book, those motifs are not aimed at fairness. They are shaped by natural selection operating *under* a fairness-like constraint. Evolution explores, recombines, and prunes designs that happen to function within the guardrails set by the world’s laws. If the Law of Fairness (LoF) is one such guardrail—an end-of-life neutrality constraint on experienced affect—then natural selection operates inside that boundary the way it does inside gravity or thermodynamics: opportunistically, locally, and without purpose.

This part lays out that story in plain language and tractable biology. We show how a fairness constraint can be compatible with, and even explanatory for, familiar evolutionary regularities without smuggling in teleology. The genome does not “try” to be fair; it simply cannot stably encode strategies whose realized streams of experience would, in aggregate, break hedonic neutrality at closure. Designs that bump into that wall either (a) never stabilize because they create uncompensable experiential debts, or (b) are canalized by development and social scaffolding so that the organism’s day-to-day options remain within admissible bounds. What we observe, then, are phenotypes that look as if they anticipate balance—not because they foresee an end state, but because only those phenotypes persist in populations where the guardrail is always on.

What this Part will do for you:

- Clarify constraint versus purpose. A constraint removes paths that would otherwise be reachable and thereby reshapes statistics of what survives. No foresight is implied. Selection “finds” paths through a rugged landscape by restricting which choices are survivable. If LoF holds at the level of experienced streams, then genotypes whose behavioral repertoires repeatedly keep lives inside admissible sets will leave more copies than genotypes whose repertoires afford rigid, brittle, or fragile counterweights. This differential persistence is a statistical consequence of admissibility constraints, not teleology.
- Identify design motifs. Selection tends to preserve under LoF these recurrent features: opponent processes that buffer extremes; flexible control hierarchies and re-weightable goals; widened choice sets; and social scaffolds that make repair easier after shocks. Executive control and social support bias the future toward admissible moves. Sleep and dream architecture support consolidation

and affective recombination. Low-cost adjustments reset the next day's starting point, making neutrality feasible on longer horizons. Social anchors, kin selection, reciprocal altruism, and rituals of reconciliation create external buffers. Individuals embedded in such structures probabilistically gain more opportunities to counterbalance harm with repair. Analgesia and palliative capacities, attention gating, and cultural pain-management practices reduce debt near closure and are not erased by selection.

- Confront edge cases directly. Could selection not simply optimize reproductive success and ignore fairness entirely? Yes—if doing so never produced uncompensable experiential drifts. But strategies that *systematically* induce extreme, unrepairable hedonic debts (for oneself or tightly coupled others) correlate with developmental collapse, social retaliation, desertion by coalitions, or physiological breakdowns—all empirically familiar filters. What remains in stable populations is not fairness-seeking, but *fairness-respecting*. LoF functions like a conservation law that everything else must work around.
- Make the ideas testable. If LoF acts as a global boundary condition, we should see convergent regularities across lineages and ecologies *over and above* what standard viability models predict. *For instance:* conserved symmetry in affective range (across species with very different bodies, the reachable distribution of day-to-day affect should be bounded and roughly symmetric over time, after controlling for ecology and life history—not because organisms “seek balance,” but because designs without such bounds fail to persist); crisis repertoires (when horizons shrink—illness, seasonal scarcity, terminal phases—organisms should show repertoire compression toward relief and repair, even when short-term reproduction might argue otherwise, because those are the moves that keep closure feasible); developmental canalization of buffers (traits for sleep, social soothing, and pain control should be strongly canalized and resilient to perturbation, reflecting selection for keeping streams admissible—attempts to knock out these systems often reduce survivability or produce dramatic dysregulation); cross-species “assistance signals” (vocalizations, gestures, and behaviors that elicit care from conspecifics—e.g. infant-like distress cues—should be widespread, because they reliably expand the menu of reparative actions after harm). None of these patterns invoke purpose or altruistic foresight. They are simply what falls out when blind evolutionary search operates under a wide-scope constraint.

In the chapters that follow, we spell out the constraints that genomes cannot break if experienced neutrality is guaranteed at closure, and show how social and cultural practices act as externalized buffer layers. We derive concrete predictions that could confirm or disconfirm this constraint-based story, and explain why LoF can be fitness-neutral in principle—evolution need not “like” or “dislike” it—while still being constraint-binding in practice.

Chapters in this Part:

- **Chapter 17 — Natural Selection Meets a Law** - draws the bright line between selection’s freedom to optimize and the non-negotiables imposed by LoF. We outline concrete evolutionary “no-go zones” in phenotype space and give everyday examples of how organisms skirt them without invoking any cosmic purpose.
- **Chapter 18 — If Life Is a Game** - turns to social and cultural expressions of the law. If many lives are coupled, one agent’s menu changes can shrink another’s. We show how norms, laws, and rituals function as population-level mechanisms to keep each individual’s admissible set open enough to avoid systemic violations (a perspective that also sets up a simulation analogy for life).

Evolutionary game theory demonstrates that reciprocal altruism – essentially a fairness strategy of “I help you, you help me” – tends to emerge as an evolutionarily stable solution among self-interested agents (Trivers, 1971; Axelrod, 1984). In other words, even in a competitive arena, balanced give-and-take often wins out, echoing the idea that designs respecting fairness constraints have survival advantages.

If the earlier parts of this book asked, “*What would we expect to see inside a mind if a fairness law holds?*”, this part asks, “*What kinds of minds and societies would evolution leave standing after a billion iterations of trial, error, and deletion under that law?*” The answer is not uplifting or grim. It is structural: designs that keep options open for repair, that moderate extremes, and that recruit help when horizons shrink will be the designs we find—again and again—because there are no lasting alternatives.

Where we go next:

We now move to Chapter 17, which examines nature’s guardrails. It shows how natural selection, unguided and pragmatic, still favors designs that respect LoF’s constraints — and what evolutionary patterns would confirm or refute the law’s influence.

Chapter 17 — Natural Selection Meets a Law

As François Jacob put it, “*Evolution is a tinkerer, not an engineer.*” It tries things blindly, keeps what works well enough, and discards what breaks. Usually, “what works” is defined solely by viability and reproduction. In this book we add a second, orthogonal filter that acts at the level of experienced lives: the Law of Fairness (LoF), which guarantees hedonic neutrality at the death of mind. The key claim of this chapter is simple:

If LoF is real, evolution still looks like evolution—but the menu of viable phenotypes is pruned by an extra, global guardrail that genomes cannot cross. What persists across time are organisms and social ecologies whose day-to-day repertoires keep experiential ledgers compensable.

This is not teleology. There is no inner mission to “seek fairness.” Instead, LoF functions like a conservation law or boundary condition. Phenotypes that systematically force large, unrepairable hedonic debts—whether for the individual or for tightly coupled others—are less likely to stabilize. They fail through developmental fragility, social retaliation, coalition loss, physiological collapse, or the sheer improbability of reaching neutral closure. Conversely, designs that incidentally enable repair, buffering, and horizon-sensitive restraint are more likely to persist. Natural selection thus discovers fairness-respecting designs *without ever aiming at fairness.*

What “Meeting a Law” Means. To “meet” gravity, wings must generate lift; to “meet” thermodynamics, metabolism must export entropy. To meet LoF, neural and social control systems must keep an agent’s admissible set of thoughts and actions rich enough, across changing horizons, to allow the lifetime hedonic ledger to come back toward zero at closure. This requirement pushes evolution to conserve and refine certain motifs that:

- Limit runaway extremes (opponent processes, refractory periods, tolerance mechanisms).
- Add low-cost counterweights (sleep/dream architecture, play, ritual).
- Recruit help on demand (attachment, empathy, reconciliation and repair practices).
- Widen option sets near closure (analgesia, caregiving, social “absolution” behaviors).

- Permit flexible reframing (prefrontal cognitive control, perspective shifts, cultural meaning-making).

These motifs are familiar because they are everywhere. Under LoF, their ubiquity is not accidental: they are exactly the traits that keep experiential “books” auditable.

Constraint, Not Purpose. LoF is a constraint, not a goal or guiding purpose. Constraints do not steer each step; they remove entire classes of trajectories from ever becoming common. Imagine an evolutionary landscape of behavioral strategies. Without LoF, many ravines might lead to fitness peaks, including some that exact catastrophic experiential costs along the way. With LoF, any ravine whose typical journey ends in an uncompensable debt is effectively walled off. Populations still climb peaks—but only along paths that keep ledgers repairable. The result is the same blind “tinkering” we know from biology, now occurring within a slightly smaller, more structured search space.

A Working Heuristic for Evolution Under LoF. We can think of evolution under LoF in three layers, each shaped by ordinary selection yet bounded by the fairness law:

- Layer 1: Body and base homeostasis. Keep tissues in range; avoid nociceptive (pain) overload; maintain energy balance. These systems minimize biological catastrophe, indirectly curbing extreme negative affect.
- Layer 2: Brain control motifs. Balance drives; implement opponent processes; enable flexible re-weighting of goals and enlist social aid when needed. These systems minimize behavioral catastrophe and keep counterbalance options open.
- Layer 3: Social scaffolding. Families, coalitions, rituals of repair, institutions of care and justice. These systems minimize interactive catastrophe, providing external buffers when individual options narrow.

LoF doesn’t *build* these layers—selection does. But LoF helps explain why all three are so stubbornly conserved, and why populations that erode one layer (e.g. chronic sleep loss or persistent social fragmentation) show escalating signs of experiential imbalance and collapse.

This aligns with the concept of allostatic load in physiology: the cumulative “wear and tear” of chronic stress responses eventually breaks the system (McEwen & Stellar, 1993). In short, organisms that cannot intermittently rebalance – shedding their stress debt – suffer mounting dysfunction, supporting the idea that uncorrected imbalance is unsustainable in the long run.

Everyday Intuitions, Reframed. Many ordinary observations about life make new sense through the lens of LoF:

- “Why do intense highs fade?” Opponent processes and tolerance often feel like “nature spoiling the party.” Under LoF they also serve as safeguards against ledgers drifting irretrievably positive (which would eventually crash negative).
- “Why do we reconcile at the end?” Deathbeds so often bring apologies, forgiveness, last visits. Culture calls it meaning; LoF sees repertoire compression under short horizons—repair options become easier to access because they keep neutrality feasible when time is almost up.
- “Why can cruelty persist?” LoF is not a moralizer. Great harm can occur, sometimes on large scales. The claim is that lives running intensely exploitative or cruel strategies cannot stably avoid counterweights—physiological, psychological, or social—that pull the lifetime ledger back toward zero.

What We Expect to See in Biology. Across species (after controlling for ecology and life history), we expect certain signatures if LoF holds:

- Bounded affective range. The reachable day-to-day affective state should be constrained to a window that is roughly symmetric and self-correcting over time. This would not be because organisms seek balance, but because designs without such bounds fail to persist in the long run.
- Crisis repertoires. When horizons shrink (e.g. seasonal scarcity or terminal illness), behavior should shift toward relief and repair—even at a short-term reproductive cost—because those actions keep neutral closure feasible. We should see creatures “pulling in” and tending to wounds (literal or figurative) as the end approaches.
- Conserved buffers. Sleep architecture, infant-care behaviors, social-soothing signals, analgesic pathways—these traits are deeply conserved and resilient. Attempts to abolish them often result in reduced survivability or dramatic dysregulation. This suggests selection fiercely protects these compensatory channels.
- Assistance signals. Cross-species cues that elicit care (for example, infant-like facial schemas or distress calls) are widespread because they reliably expand an individual’s admissible set after harm. Even different species sometimes respond to each other’s distress signals; evolution kept those signals because they widen options for recovery.

What We Do Not Need to Assume. Notably, LoF does *not* require us to assume any of the following:

- No foresight. Organisms need not compute a lifetime ledger or know their future; short-horizon adjustments emerge without any long-term planning or awareness of an “end state.”
- No moral desert. “Fairness” in this context is a neutral global accounting, not cosmic justice. There is no implication that individuals *deserve* their outcomes in a moral sense—only that ledgers tend toward balance by the end.
- No hidden miracles. All proposed mechanisms are standard biological and social control motifs, simply reinterpreted as buffer machinery. We invoke no mysterious forces beyond known biology.

Testable Predictions (Preview). Later chapters will lay out full experimental designs, but here are a few short-form predictions geared to an evolutionary context:

- Convergent buffering. Distant lineages should independently evolve functionally similar repair affordances (e.g. distinct species converging on sleep phases or social reconciliation rituals) whenever social complexity and lifespans pass certain thresholds. (In short, similar problems get solved in similar ways across evolution if LoF holds.)
- Horizon compression. In longitudinal wildlife studies, end-of-life phases should show repertoire narrowing toward relief and reunion more than matched midlife controls (controlling for frailty and other factors). As the horizon shrinks, behavior compresses around making amends and seeking comfort.
- Canalization of care. Genetic or developmental perturbations that erode basic care-elicitation or care-giving behaviors should carry outsized fitness penalties relative to neutral traits. In other words, breaking the social-soothing or caregiving machinery should be *extremely* costly, reflecting its key role in maintaining admissible options for the organism.
- Population-level tilt. Societies that institutionalize repair (for example, restorative justice programs, hospice practices, conflict-cooling rituals) should show lower variance in late-life affect, after matching for factors like wealth and healthcare. The idea is that a strong social layer bolsters admissible sets near closure, leading to more consistent neutral outcomes across individuals.

Why This Matters for the Rest of the Book. Earlier parts of the book worked “inside the mind,” deriving signatures of a regulatory Queue System that prunes thought-action

menus in ways that preserve hedonic compensability. Here, we step back to the evolutionary timescale and ask: *would blind selection, if bounded by LoF, tend to conserve the very motifs that give the Queue System (QS) enough room to work?* Our answer is yes. Evolution gives QS its levers—sleep, sociality, opponent dynamics, flexible control—and culture then reinforces those levers through norms and institutions.

What you'll get from this Chapter:

- Constraints the genome can't break: Understand why, if LoF holds, natural selection can't produce organisms whose unified streams run permanently net-positive or net-negative in felt experience. We clarify what counts as a “no-go” zone for evolved life.
- Why control systems resemble QS: Recognize how evolved regulators of behavior and emotion (hunger, pain, pleasure, etc.) act like guardrails that prevent runaway states, and in many ways resemble a Queue System rather than any purpose-driven design.
- Cross-species predictions: Anticipate what *other* species might show if LoF applies broadly. For example, social mammals might exhibit end-of-life calming or horizon-driven behavior changes, whereas very simple creatures might not—suggesting comparative tests to distinguish LoF-driven regulation from mere fitness maintenance.
- Fitness-neutral vs. constraint-binding (Research Note): Appreciate experimental designs that hold survival/reproductive fitness constant while manipulating emotional balance. These help us avoid conflating fairness effects with simple viability, clarifying how LoF could be *fitness-neutral* in theory yet still impose hard constraints on evolutionary trajectories.
- Fail patterns with systematic imbalance: Identify what evidence would spell trouble for LoF. If we ever found a lineage, species, or engineered organism that reliably accumulates net suffering or net pleasure across life without counterbalancing forces *and* without incurring survival penalties that enforce balance, LoF would be in serious doubt.

Subsections in this Chapter

- **17.1 Constraints the Genome Can't Break** - Spells out the phenotype “no-go zones” that evolution cannot stabilize if LoF holds: organisms whose unified streams drift permanently net-positive or net-negative; architectures that

foreclose counterweights; and regulatory designs that allow unbounded hedonic accumulation. We translate those impossibilities into concrete selection pressures and trait-level predictions to look for in real lineages.

- **17.2 Why Control Systems Resemble QS** - Shows why durable biological controllers converge on QS-like motifs: bounded highs/lows, horizon-weighted control gains, supervisory switching under stress, and cheap “offline” rehearsals (e.g., during sleep). Includes a field/lab checklist for testing these signatures and a bottom-line argument that such motifs are expected if a compensability-preserving constraint is in play.
- **17.3 Cultural Echoes: Karma, Justice, Penance** - Surveys cross-cultural “ledger” technologies—karma doctrines, confession/penance, restitution—that track moral debt/credit, scaffold repair, and intensify balancing moves near horizons (e.g., end-of-life rites). Not offered as proof of LoF, but as convergent social designs mirroring QS features, with testable predictions (e.g., repair-metaphor density vs. societal stability).
- **17.4 Cross-Species Predictions** - Outlines comparative predictions: social mammals should show horizon-sensitive calming and end-of-life compression, while very simple agents may not. Proposes targeted cross-species assays (e.g., variance in affect near terminal horizons, presence/absence of “offline” counterweighting) to distinguish LoF-consistent regulation from mere fitness maintenance.
- **17.5 Research Notes: Fitness-Neutral, Constraint-Binding** - Details experiment designs that hold survival fitness approximately constant while manipulating balance mechanisms (or vice-versa), so we can test whether neutrality arises from a constraint rather than as a by-product of viability. Lays out preregistered comparisons and metrics for adjudicating the two.
- **17.6 Fail Patterns: Species with Systematic Imbalance** - Specifies decisive falsifiers: any lineage or engineered strain that reliably accumulates net pleasure or net suffering by terminal time—without counterweights and without survival penalties enforcing balance—would directly challenge LoF. Lists observable footprints such a species would leave in longitudinal affect and behavior.

Where we go next:

We begin by marking the evolutionary “no-go zones.” In 17.1, we enumerate the designs evolution cannot stabilize if LoF is true and show how those impossibilities become selection pressures that favor QS-shaped regulation. This sets up 17.2–17.6, where we trace the expected motifs in real control systems, their cultural echoes, comparative predictions, constraint-binding tests, and the Fail patterns that would overturn the law.

17.1 Constraints the Genome Can't Break

Natural selection is permissive about *how* organisms win, but it is ruthless about designs that cannot keep winning long enough to leave descendants. If the Law of Fairness is a boundary condition on experienced lives—guaranteeing that each stream closes near hedonic neutrality—then certain phenotypic “strategies” will be filtered out, not because they violate morality, but because they systematically strand agents with uncompensable ledgers. This section catalogs those evolutionary no-go zones, translates them into recognizable biological motifs, and extracts testable predictions.

This view is consistent with established evolutionary models. For example, Hamilton’s rule (Hamilton, 1964) and Trivers’ reciprocal altruism (Trivers, 1971) both imply that organisms evolve to avoid uncompensated costs to kin or cooperative partners. In effect, genes will favor strategies that preserve balance in social exchanges, since alleles causing net disadvantage (without compensation) are weeded out. In other words, classical kin-selection and reciprocity theory already predict a bias toward fairness, aligning with LoF’s basic premise.

Comparative studies show that many social animals display inequity aversion, refusing or protesting unfair splits of reward. Such ancient responses likely evolved to preserve cooperation in groups. This cross-species evidence indicates that inequity sensitivity is widespread; whether this reflects a fairness constraint or standard cooperative dynamics is an empirical question rather than a settled conclusion.

Analyses of human and animal behavior suggest that stronger fairness preferences may co-evolve with reliance on cooperation, consistent with such mechanisms carrying adaptive value in cooperative contexts.

17.1.1 Three classes of forbidden design

We can group the “unviable under LoF” designs into three broad classes:

Runaway amplifiers of uncountered pain or pleasure.

Definition: Mechanisms that allow sustained, one-sided drift of experienced affect without embedding low-cost counterweights or refractory gates.

Examples: Pain circuitry that scales linearly with injury without any analgesic ceiling; reward loops with uncapped positive feedback and no tolerance or satiety mechanism.

Why forbidden under LoF: Such systems would make it highly improbable to re-enter a compensable range within typical lifetimes, raising the probability that a life ends with a large net surplus or deficit (a broken ledger).

Ledger-blind commitment engines.

Definition: Traits that lock organisms into long sequences of behavior carrying high risk of massive hedonic deficits, with minimal branching opportunities to repair or exit.

Examples: Irreversible “vendetta” modules (instincts that enforce prolonged retaliation); compulsions that escalate in cost while narrowing options; extreme pair-bonding that cannot be reconfigured after catastrophic loss.

Why forbidden: As horizons stochastically shorten (through disease, predation, accidents), these one-track engines leave too few admissible exits to regain neutral closure. They court experiences that cannot be counterbalanced in time.

Isolation from external buffers.

Definition: Designs that drastically suppress attachment bids, social soothing, or coalition repair—beyond rare contexts—thereby shrinking the pool of admissible counterweights when internal regulation fails.

Examples: Obligate solitary life histories in species with long juvenile dependency (no social fallback when stress hits); chronic suppression of care-eliciting signals even under extreme distress.

Why forbidden: When internal regulation falters, only social scaffolding can rapidly widen the menu. Phenotypes that chronically disable this channel become extremely fragile near closure, as they cannot recruit last-minute relief.

Heuristic: If a design makes neutral closure *unlikely* for typical lives across typical shocks, it struggles to persist. The genomes that remain are those that implement bounds, branches, and buffers.

17.1.2 What the surviving motifs look like

Across taxa, we therefore expect to see canalized (developmentally stabilized) motifs that keep experiential ledgers compensable. In broad strokes, they fall into three categories—Bounds, Branches, and Buffers:

Bounds:

Opponent processes and tolerance (e.g., antinociception after injury; hedonic adaptation after windfalls) that counteract prolonged highs or lows. *Satiety and satiation mechanisms* that cap drive-driven sequences (for food, sex, etc.), preventing endless pursuit or pain. *Habituation and sensory gain control* that prevent persistent overload from unchanging stimuli.

Branches: *Flexible action selection with reversal options* and “exit ramps” (e.g., appeasement displays, conflict-cooling gestures, reconciliation rituals) that allow an organism to escape escalating harm or to pivot to repair mode. *Cognitive reframing capacities* that reopen or reconfigure goals when contingencies change (e.g., prefrontal control enabling one to drop unattainable goals, or model-switching to find new solutions).

Buffers: *Sleep architecture* (alternating NREM/REM cycles) that provides low-cost physiological and emotional recalibration—essentially nightly hedonic counterweights. *Attachment systems* that reliably elicit care from conspecifics (infant crying and “cute” features, distress vocalizations, social grooming invitations). *Community practices* (in humans and some social species) that institutionalize repair (pair-bond maintenance, coalition reconciliation behaviors, communal caregiving, hospice-like end-of-life care).

These traits were not built “for” fairness per se; they were retained because phenotypes lacking them are less likely to navigate a stochastic life course with admissible menus intact. In other words, organisms missing bounds, branches, or buffers would more often get stuck with irreparable deficits or surpluses, and thus be pruned out over evolutionary time.

17.1.3 Formalizing the constraint

Let $L(T)$ denote the running hedonic ledger for a conscious stream, and let T be the (possibly stochastic) time of closure (the death of mind). Operationally, LoF requires:

$\Pr(|L(T)| \leq K \mid \text{typical shocks, typical supports}) \geq 1 - \varepsilon$, for some finite bound K (in HCU) and small ε , where probability is taken over the stochastic life trajectory induced by ordinary perturbations—this condition applies at the level of whole lives, not momentary states.

A phenotypic design class \mathcal{D} is admissible if, for the ecological distribution of life trajectories it induces, the above inequality holds.

Designs for which ordinary perturbations produce $\Pr(|L(T)| \leq K) \ll 1 - \varepsilon$ are selected against over evolutionary time because they routinely strand conscious streams with uncompensable ledgers.

The canalized motifs listed above are precisely those that keep $\mathcal{A}(t)$ (the admissible option set) non-empty under short horizons, thereby raising the probability that closure lands within the $[-K, K]$ band.

17.1.4 Developmental canalization and repair affordances

Under LoF, we expect early-developing circuits to lock in capacities that are unusually protective of compensability:

Attachment first: Neonatal bonding, separation distress, and caregiver soothing behaviors develop early and are hard to extinguish. *Prediction:* Perturbing these modules (e.g., by preventing normal bonding) produces not just social deficits but amplified late-life variance in affect (wider emotional swings), even when basic needs (nutrition, safety) are met. In short, early attachment disruptions would leave lifetime ledgers wobblier, consistent with a built-in buffer being removed.

Sleep robustness: Sleep rebounds after deprivation, and REM sleep proportion adjusts to stress. *Prediction:* Species with longer, more complex horizons (long lifespans, high social complexity) should show stronger homeostatic recovery of REM after stressors, and loss of REM should have disproportionate effects on ledger variance. (In simple terms: the more an animal needs to keep experience balanced over decades, the more its biology “fights for” its REM/dream counterweights.)

Play and rehearsal: Juveniles of many social species play-fight and role-switch in seemingly non-productive ways. *Prediction:* The extent of juvenile play correlates positively with the availability of repair scripts in adulthood and with narrower affect variance near life stressors. Play may be evolution’s way of rehearsing counterbalances (e.g., learning to reconcile after conflict) before stakes get high.

17.1.5 Social ecology as an externalized regulator

Genes cannot guarantee the *right* experiences in every contingency, but they can make social help cheap and likely:

Signals that scale with need: Many species have distress calls that intensify with threat and reliably elicit aid rather than exploitation. *Prediction:* In species where distress signals are sometimes exploited by others, those signals evolve to be costly to fake (the handicap principle) or otherwise honest, preserving their compensatory function. (If being vulnerable reliably invites help, not attack, it widens admissible moves under duress.)

Reconciliation protocols: After aggression, many primates and other mammals engage in reconciliation and consolation behaviors. *Prediction:* Groups with clear post-conflict reconciliation scripts show faster returns to baseline affect and less late-life drift in well-being, controlling for dominance hierarchy and other factors. In effect, lineages that “fix” bad blood quickly will keep individual ledgers more balanced.

Terminal support: In humans (and perhaps some social mammals), hospice care and communal end-of-life rites emerge when horizons shrink. *Prediction:* Access to terminal support (from caretakers or community) compresses the distribution of end-of-life HCI (Hedonic Composite Index) for individuals, reducing both extreme highs and lows. This effect should hold independent of wealth or general healthcare—a cultural buffer at work.

17.1.6 How LoF filters extremes without moralizing

Consider two stylized evolutionary strategies:

Predatory maximizer: This organism gains large, repeated pleasure by exploiting conspecifics and invests little in repair or care. Short-term fitness may be high, but social retaliation, coalition exclusion, and internal dysregulation soon narrow its future admissible sets. Over a lifetime, horizons can shorten abruptly (illness, injury), making neutral closure improbable especially because the strategy has corroded the very social buffers that might rescue it. *Prediction:* Such exploitative strategies remain rare and short-lived as stable life histories. They might pop up as transient tactics (when conditions allow a quick win), but you won't find entire lineages that sustain "cheater" strategies without compensatory collapse.

Buffer-builder: This organism gains moderate pleasure, accepts moderate pain, and actively invests in repair channels (e.g., mutual grooming, nested social bonds, periodic rest). It may sacrifice some short-term gains, but it maintains wider menus of action under stress, especially near closure, thereby keeping a high probability of neutral endings.

Prediction: Buffer-building motifs become common in successful lineages; cultural scaffolds often evolve to amplify them. In social species, we expect to see norms that reward moderate strategies and support their repair behaviors (since those groups thrive collectively).

17.1.7 Cross-species probes

We can generate specific comparative predictions by looking at particular kinds of species:

Cetaceans and elephants: These have long lifespans, rich social networks (alloparenting, post-reproductive helpers), and even signs of grieving rituals. *Prediction:* They will exhibit strong social repair behaviors and perhaps ritualized mourning that stabilizes ledgers after a loss. Look for horizon-sensitive behaviors in aged individuals that prioritize affiliation and reconciliation over exploration or competition.

Corvids and parrots: These birds have high cognitive flexibility and long lives relative to body size. *Prediction:* They will have branch-rich repertoires—many options to avoid or exit bad situations. Specifically, test whether conflict-cooling gestures (like allopreening or gift-giving) and food-sharing increase as individuals age or after major stressors, indicating built-in repair routines.

Cephalopods: Octopus and cuttlefish are short-lived but surprisingly intelligent, with rapid senescence. *Prediction:* Because of their short natural horizons, we expect minimal late-life *behavioral* compression (they simply don't live long enough for protracted “retirement” phases). However, they should still show bounded affective ranges (no endless terror or mania) and possibly sleep-like states that act as low-cost buffers even in their brief lives.

17.1.8 Human cultural inventions as multiplicative buffers

Cultural practices that persist across centuries often instantiate repair affordances—essentially externalized QS interventions:

Ritual confession, apology, and forgiveness create structured exit ramps after transgression, allowing wrongdoers to rejoin the group and emotional balances to reset.

Rest days, sabbaths, festivals enforce periodic down-regulation (rest, social bonding, celebration) that prevents hedonic ledgers from drifting too far during continuous labor or stress. In essence, culture mandates counterweights (rest or joy) after work and struggle.

Hospice and palliative care widen admissible options near closure, enabling peace with unresolved projects and relationships. These practices give people chances to reconcile and find comfort at end-of-life, rather than leaving them in unchecked despair.

Prediction: Societies that normalize these practices (regular atonement rituals, rest periods, end-of-life care) will exhibit lower end-of-life HCl variance and fewer extreme outliers, after adjusting for medical care and material conditions. In short, cultures that institutionalize counterbalancing will show more consistent neutrality at life's end.

17.1.9 Where we go next:

In Section 17.2, we explore why evolved control systems mirror the Queue System’s logic. We will see how biological mechanisms—from neural circuits to behavior—exhibit patterns of horizon-sensitive regulation, reinforcing the parallels between natural selection’s outcomes and the fairness constraint.

17.2 Why Control Systems Resemble QS

Biology rarely steers by brute force; it steers by constraints—saturating gains, refractory periods, competing controllers, layered feedback loops. When we view organisms through that lens, familiar control architectures begin to look like local implementations of the Queue System (QS). This is not because cells and circuits *know* about fairness, but because the only control systems that survive across many shocks are those that keep admissible menus open and compensation feasible, especially as horizons shrink.

Indeed, fairness-like constraints appear even in basic social animals. For example, capuchin monkeys refuse unequal exchanges: Brosnan & de Waal (2003) found monkeys would reject a cucumber slice if another monkey got a more desirable grape for the same task. Such inequity aversion suggests an evolutionarily ancient bias against uncompensated disadvantage, consistent with our idea that organisms evolve control motifs enforcing balance.

17.2.1 The control-theory template

A minimal control system (regulator) has: (i) a plant (the organism–environment dynamics to be controlled), (ii) sensors (what's monitored), (iii) a controller (the policy or decision unit), and (iv) actuators (the means by which the controller affects the world). Robust controllers augment this with additional features:

Saturation and bounds – hard limits that prevent runaway behavior.

Integral terms – slow accumulators that correct long-standing error.

Adaptive gains – the ability to change the strength of control as conditions vary.

Supervisory layers – a meta-controller that can switch policies when one is failing or inappropriate.

QS maps onto these pieces functionally:

The ledger $L(t)$ behaves like an integral of hedonic error (an accumulation of positive or negative affect over time).

The horizon H_t acts as a supervisory context that scales gains (analogous to the shadow price λ_t from earlier chapters).

The admissible set $\mathcal{A}(t)$ is the bounded actuator space—only options within the constraint are allowed to be enacted.

The QS residual functions like a meta-signal that biases policy switches toward repair when the ledger is way off balance and horizons are short.

17.2.2 From homeostasis to homeodynamics

Classic homeostasis tries to hold internal variables near fixed set-points (temperature, blood glucose, etc.). But organisms live in stochastic worlds; fixed set-points can be fragile. Modern physiology looks more like homeodynamics: a moving target with context-sensitive ranges and multi-objective trade-offs (e.g., immune vigor vs. oxidative stress). QS adds one more dimension to this: keep experience compensable by preventing long bouts of one-signed (all-positive or all-negative) affect without counterweights. The control motifs that realize this include:

Opponent processes (bounds): Any prolonged high or low affect tends to recruit an opposing process (e.g., analgesia after intense pain, tolerance after a sustained reward). This puts brakes on runaway pleasure or pain.

Allostatic resets (supervision): Under chronic load, systems may adopt new baselines (allostasis), but always preserve *exit ramps*—behaviors that can reopen compensability. Examples include sleep rebound after prolonged stress, or social seeking and “time-outs” after trauma. These are supervisory adjustments to keep the organism viable.

Gain scheduling by horizon: As H_t shortens (the future time available shrinks), the controller *increases* the gain on repair-enabling behaviors and *decreases* the gain on risky indulgences. In life, we see this as people becoming more cautious and reconciliation-focused when they sense time running out.

Prediction (P1): In tasks that experimentally manipulate perceived horizon (e.g., a deadline or a mortality salience scenario), we should see controller gain scheduling: stronger inhibitory thresholds for low- Φ options and facilitated selection of high- Φ (repair or relief) options, even when immediate rewards are held constant. In other words, when subjects feel the “end” coming, their internal controllers should pivot toward balance-preserving choices.

17.2.3 Supervisory control in cortex

At the neural level, QS-like supervision shows up as *layered controllers* in the brain. Different regions play specialized roles:

vmPFC/OFC: tracks multi-attribute value and policy *viability* (i.e. it represents not just reward, but whether a plan can be pursued without downstream collapse).

ACC: monitors control cost and conflict, ramping up activity when a continuing course of action looks increasingly uncompensable.

rIFG: implements a context-sensitive “brake,” raising the threshold on actions that are deemed non-admissible.

Insula: integrates interoceptive load (internal bodily stress signals), tipping the system toward relief/repair behaviors when physiological debt accumulates.

This is standard control logic wearing a neural mask: these are functional attributions, not one-to-one anatomical claims. These components closely mirror what a QS controller would require.

Prediction (P2): When two choices are matched on immediate reward and risk, the more repair-enabling option will show an extra vmPFC signal and *lower* ACC conflict *only* under conditions where $|L(t)|$ is large or H_t is short. In other words, a significant interaction between ledger imbalance and horizon on brain signals (a $\Phi \times H$ effect) is the signature of QS-like supervision. If the brain is truly optimizing just expected utility, we wouldn’t see that specific interaction.

17.2.4 The admissible actuator space

Mechanical controllers have saturating actuators: you can’t push a motor beyond its torque limit. Minds, analogously, have *realizable limits* on actions: some options never even come to mind or never reach the brink of execution if they would threaten ledger compensability. This isn’t magic or fate; it’s the controller failing to supply sufficient motive force for those actions. In practice:

The thinkable set (what you can conceive of doing) remains broad.

The selectable set (what you’d seriously consider after weighing costs) narrows under high conflict or risk.

The admissible set (what actions actually *trigger* motor plans) narrows further under QS influence—specifically when $H \downarrow$ or $|L|$ is extreme.

Prediction (P3): As natural horizons approach (end of a semester, impending surgery, advanced age), people’s language and behavior corpora will show menu shrinkage and a tilt toward repair: more reconciliations, completions of unfinished business, consolidations of social ties; fewer novelty-seeking or high-variance gambles. This holds even after controlling for age, health, and culture. In short, near closure, the admissible set contracts to what keeps things balanced.

17.2.5 Integral control and the ledger

In control theory, an *integral term* in a controller accumulates error over time to correct persistent biases. But too much integral wind-up causes overshoot or instability. The hedonic ledger $L(T)$ is exactly such an integral of momentary affect $F(t)$:

$L(t) = \int_0^t F(\tau) d\tau$, with QS enforcing that $|L(T)| \leq K$ with probability at least $1 - \varepsilon$ (as defined earlier). Whenever $|L(t)|$ grows large in magnitude, QS increases λ_t (the shadow price of further one-sided affect), re-weighting the organism's policy toward counterweights. In effect, QS acts as an anti-wind-up mechanism: it prevents the integral from growing so unchecked that neutral closure becomes improbable.

Prediction (P4): Individuals with impaired “anti-wind-up” circuitry—say, lesions or disconnections affecting ACC–vmPFC coupling—will show fatter tails in lifetime well-being (HCl) measures and less compression of affect as they approach end-of-life, compared to matched controls. They will have more difficulty reigning in extreme cumulative emotions, consistent with a broken integrator clamp.

17.2.6 Multi-objective control: survival and compensability

Brains juggle multiple objectives: survival probability, resource acquisition, mating success, social rank, and so on. QS asserts an additional, orthogonal constraint: keep the lifetime affect integral near zero. In a multi-objective controller, that can be formalized as either a lexicographic priority or a penalized optimization:

Lexicographic: Maximize viability subject to $\Pr(\text{neutral closure}) \geq 1 - \varepsilon$. (In words, never accept a strategy that makes fairness-violation likely, no matter how fit it otherwise is.)

Penalized: Maximize viability minus λ_t times the expected uncompensability. (The controller pays a “cost” for policies that drive the ledger too far, especially as t approaches T .)

Either way, as horizons shrink, $\lambda_t \uparrow$ and the compensability constraint binds harder. That is why “endgame” behavior (in humans and other animals) shows robust repair motifs across cultures and contexts—things like settling debts, making amends, or retreating to safety when time is short.

Prediction (P5): In agent-based simulations where a “compensability penalty” is applied only at the lifetime level (not each step), the agents that evolve will nonetheless exhibit local signatures of QS-like behavior—opponent-process dynamics, spontaneous repair rituals, menu narrowing near horizon—that agents trained on pure stepwise rewards will not. In other words, even if the fairness constraint is global, evolution will bake its influence into everyday behavior.

17.2.7 Dreams as a scheduler’s trick

Control engineers reduce risk and cost by running stressful tests offline (in simulators) and pre-computing compensatory moves in advance. REM sleep dreams look suspiciously like that: a low-cost “sandbox” where negative scenarios are rehearsed (threat simulation yielding negative prediction errors) and positive re-interpretations are tried (wish-fulfillment and social reconciliation narratives) without incurring full metabolic or social cost. In this framework, we predict that REM intensity boosts after days of high negative load; that dream affect can invert relative to the prior day’s experiences (bad days yielding more positively toned dreams, and vice versa); and that dream content tracks salient personal debts and goals. This would be the controller using an internal simulator to keep the real-world admissible set wide.

Prediction (P6): After high-load days (low well-being, negative HCl readings), REM sleep proportion and *counter-valenced* dream content increase, and the next-day admissible menu widens (more repair-oriented choices, less perseveration on hopeless options), beyond what would be explained by sleep duration alone. In short, we expect to see that dreams actively contribute to resetting the ledger.

17.2.8 Why this resemblance matters

The resemblance between evolved control systems and our theoretical QS is not cosmetic; it’s functional. It matters because it provides a unifying explanation and a framework for falsification:

It predicts new phenomena (e.g., specific horizon × value interactions in decision-making, late-life “menu tilt” toward repair).

It organizes known ones (opponent-process physiology, reconciliation rituals, REM rebound after stress) under a single constraint-based story rather than treating each as an independent quirk of evolution.

It falsifies cleanly: If we can show a persistent *absence* of horizon-sensitive gain scheduling, of menu shrinkage near closure, or of ledger compression in well-measured life streams, then the QS analogy—and by extension LoF’s role—fails.

In other words, if QS were just metaphorical hand-waving, there would be no particular reason for control systems to repeatedly converge on these motifs. The fact that we *do* see bounded highs and lows, context-adaptive control gains, last-minute behavioral switches from indulgence to repair, and “offline rehearsals” (dreams) suggests that something like QS is at play. Either that, or these patterns are all coincidental outcomes of simpler objectives—which is testable.

17.2.9 Quick checklist for field and lab

To probe QS dynamics empirically, one can:

Manipulate horizon (e.g., impose time pressure or mortality salience) and test for $\Phi \times H$ interactions in vmPFC/ACC activity and in choices. Do values and conflicts shift uniquely when time is perceived as short?

Estimate person-specific $L(T)$ from hedonic data (HCl) and look for anti-wind-up signatures: does a prolonged run of one-sided affect automatically increase the weight on repair behaviors (or thoughts) subsequently?

Track actuator saturation: Are certain action-options reliably failing to reach execution threshold precisely when taking them would deepen an uncompensable debt? (E.g., does revenge not even occur to someone who is already severely down because it would make things worse?)

Probe supervisory switching: Do people flip from exploration mode to repair mode when $|L| \uparrow$ or $H \downarrow$, even if the immediate reward of continuing is high? (This could be tested with tasks where participants can either exploit a resource further or pivot to a restorative task, under different time constraints.)

Control systems that *last* tend to be QS-shaped: they bound extremes, they schedule gains by horizon, they recruit “higher authority” controllers under stress, and they rehearse counterweights cheaply (e.g. in sleep). Evolution didn’t design brains and societies *to obey fairness*; it kept those designs because life-streams that lack such features run out of admissible moves and hit dead ends. QS is the abstract blueprint; physiology (and social behavior) is the implementation. By measuring where and how that blueprint is etched into brains, behaviors, and cultures—and looking for where it *isn’t*—we can determine whether the Law of Fairness is truly governing the “plant” (the world of experience), or whether we are merely seeing the shadows of simpler evolutionary objectives.

17.2.10 Where we go next:

In 17.3, we shift from biology to culture. After examining why evolved control systems resemble a Queue System, we next explore cultural echoes of cosmic fairness—how human moral traditions (karma, penance, justice, etc.) mirror the patterns we’d expect if LoF holds.

17.3 Cultural Echoes: Karma, Justice, Penance

Cultures everywhere reach for the language of ledgers when they talk about harm and help. We “owe” apologies, “pay” for mistakes, strive to “make amends,” “clear our conscience,” “balance the scales,” and trust that “what goes around comes around.” This chapter does not claim that religions or legal codes *prove* the Law of Fairness. Rather, it shows that across time and place, humans have repeatedly invented social technologies—karma doctrines, justice systems, penance rituals—that mirror the very features LoF predicts a compensability-preserving controller (QS) would favor: (i) tracking moral debt and credit, (ii) enabling repair through structured counterweights, and (iii) intensifying balancing moves as horizons shorten (e.g. rites of confession and absolution near death). These echoes are anthropological data points: convergent solutions that keep communities—and inner lives—within the bounds of long-run hedonic neutralization.

17.3.1 Ledger metaphors are human universals

Across unrelated languages, moral discourse leans on accounting schemas: debt/credit, burden/relief, clean/unclean, the weighing of the heart, balances and scales. On LoF, this is not a semantic accident. If life-streams that avoid uncompensable extremes are more viable (socially and psychologically), then cultures that codify a running *moral ledger*—and scaffold robust avenues for repair—should persist and spread.

Prediction (C1): A cross-linguistic corpus analysis will show a high prevalence of accounting metaphors for morality (and end-of-life appraisal) in widely separated language families. Moreover, the density of repair-metaphors (words for forgiveness, restitution, reconciliation) should correlate with societal metrics of stability: specifically, lower variance in population-level well-being (measured by something like HCI distribution) and fewer stalled, multi-generational conflicts in historical records. (Intuition: cultures that talk about making things right tend to actually resolve things.)

17.3.2 Karma as a socialized compensability heuristic

Classical karma doctrines (in Hindu, Buddhist, Jain traditions) teach that actions carry consequences that *ripen* until balance is restored—sometimes across lifetimes. LoF does not require rebirth or any metaphysical mechanism, but the *functional shape* aligns: a persistent debt register for wrongdoing or merit, structured opportunities for counterweights (*dāna* (charity), *seva* (service), vows, precepts), and horizon-sensitivity (urgent spiritual practice in later life, monastic vows at turning points). In effect, karma is a cultural model that keeps track of one’s ledger and promises that if not in this life, then eventually, accounts must balance.

Prediction (C2): Practitioners who engage regularly in structured karmic repair activities (acts of generosity, confession of faults, loving-kindness meditation, etc.) should show QS-like signatures in their lives. For example: (a) an increased selection of repair-enabling options in daily life (detected via experience sampling or wearable telemetry), and (b) a compression of hedonic variance as known “checkpoints” approach (such as before pilgrimages, fasts, or death anniversaries, when karma is thought to come due). These would suggest that the cultural practice is effectively nudging the QS.

17.3.3 Abrahamic atonement: confession, restitution, Jubilee

Judaism’s Yom Kippur (focused on *teshuvah*—return/repair), Christianity’s sacraments of confession and penance, and Islam’s concepts of *tawbah* (repentance) paired with structured restitution (e.g. *diyya* compensation, *kaffāra* expiations) all institutionalize counterweights that one can intentionally enact. These include apologizing, making material repayment, fasting, giving alms, and seeking forgiveness. The Torah’s Jubilee tradition even resets debts and frees slaves every 50 years; Islamic justice balances *qisās* (retribution as equity) with forgiveness/compensation; Catholic penance couples contrition with required acts (prayer, service to others) to restore balance.

Prediction (C3): In naturalistic studies, completion of such sacramental or juridical repair rituals should be followed by (i) acute drops in physiological stress load (e.g. increased heart-rate variability, decreased cortisol), (ii) reduced intrusive rumination about the wrongdoing, and (iii) an admissible-menu shift toward long-deferred pro-social tasks and reconciliations—effects above and beyond any placebo or expectancy. Essentially, after “making things right,” people should feel and act as if a weight was lifted (and QS would agree).

17.3.4 Scales and Nemesis: Greco–Egyptian motifs

Ancient Egypt’s concept of Ma’at culminates in the *weighing of the heart* against a feather—a literal balance-scale determining one’s fate after death. Greek mythology’s Nemesis checks excessive pride (hubris) to ensure no one escapes the balance of fortune, and Roman *aequitas* along with the blindfolded *Iustitia* personify impartial balancing of the scales of justice. Even medieval trial by ordeal can be interpreted here: these legal oaths and ordeals were costly signals aiming to restore social compensability by decisively determining guilt or innocence, followed by appropriate repair (fines, exile, service) rather than endless feuds.

Prediction (C4): Legal systems with stronger restorative provisions (e.g. formal restitution processes, mediated apologies) will produce faster decay of conflict-related arousal and shorter half-lives of retaliatory events in social networks than purely retributive systems.

This should hold when controlling for crime rates and resources devoted to enforcement. In other words, societies that “balance the scales” with restoration see conflicts truly end, rather than simmer.

17.3.5 Penance as engineered counterweight

Secular and religious penances share a structure: (i) acknowledge harm; (ii) accept a cost or hardship voluntarily; (iii) perform repair behaviors that benefit those harmed or the community; and (iv) receive reintegration or absolution. The LoF/QS reading of penance is mechanical: a penance ritual directly raises the *feasibility of compensation* Φ by creating low-friction repair channels. Notably, when horizons shrink (grave illness, impending punishment, nearing death), penance intensity reliably rises (people urgently seek to atone).

Prediction (C5): Longitudinal data will show that late-life initiations of reconciliation or charity surge even among nonreligious populations. The effect should scale with objective horizon markers (diagnosis of terminal illness, advanced age) *and* with ledger load (e.g. number of unresolved conflicts). This is consistent with a rising shadow price of unrepaid harm as per QS (see shadow price λ_t in Chapter 6). Essentially, as time runs out, even secular people start “making amends” at higher rates.

17.3.6 Carnival, fasting, pilgrimage: systemic rebalancers

Many cultural calendars incorporate periodic counterweights: fasting after feasting, charitable alms after harvest or earnings, carnival misrule (inversions of social order) before a return to sober restraint, pilgrimage journeys that renew social bonds and re-anchor values. These rituals periodically compress variance in experience and re-open admissible menus (people take on new vows, reconcile relationships, launch community projects after pilgrimages, etc.). They act like scheduled maintenance for the social ledger.

Prediction (C6): Around major calendrical rituals (e.g., Ramadan, Lent followed by Easter, Diwali, etc.), population-level well-being (HCl) time series should show *predictable variance narrowing*. We should also see increased rates of repair-coded acts (forgiveness given, donations made, outreach to estranged family) and, following the ritual period, an expansion in people’s perceived option sets (resumption of deferred responsibilities, new initiatives), compared to baseline periods.

17.3.7 Restorative justice and “making whole.”

Modern restorative-justice programs operationalize the idea of repair: victim–offender mediation, restitution plans, community service tailored to the harm done. These

programs are designed to reduce recidivism and improve victim satisfaction by inserting counterweights that standard punitive approaches can omit. They “make whole” what was damaged, aligning with the compensatory ethos. *Prediction (C7):* When two sentencing options are equivalent in terms of deterrence and incapacitation, the one with explicit repair work (like a mediated apology plus restitution/community service) will yield: (i) steeper declines in the offender’s guilt-rumination and stress markers, (ii) faster victim relief (measured by validated well-being gains), and (iii) broader social-network reconciliation (friends/family of both sides resuming contact). These are outcomes favorable to QS (ledgers balanced, relationships restored) and should appear above and beyond what a purely punitive sentence would achieve.

17.3.8 The just-world trap: guardrails against blaming victims

A great danger of “cosmic balance” narratives is victim-blaming (“They suffered, so they must have deserved it”). LoF explicitly rejects that inference. Balance is global and lifetime, not a bookkeeping of deserts per incident. The suffering of innocents is entirely compatible with LoF *only if* counterweights (often unseen to bystanders) are available or scheduled (through dreams, relationships, later-life events, etc.). We must maintain (a) ethical guardrails (see Chapter 11) and (b) strict anti-tautology in our reasoning: we do not infer that balance *must* exist just because we observe suffering. Instead, we test for the actual mechanisms—shifts in menus, horizon-based adjustments, end-of-life compression—that would indicate balance at work.

Practice note: When communicating about these ideas, always emphasize *repair opportunities*, never deserts. And in research, measure processes (menus, horizons, compressions), not mystical scorekeeping. LoF is a lawlike constraint, not a moral scorecard, and it must be presented and tested as such to avoid the just-world fallacy.

17.3.9 Secular ledgers: credit scores, reputations, apologies

Even outside overt religion, societies develop ways to track reputation capital and enforce balance. Repeated norm violations cause someone to “owe” trust or lose access to cooperation; a sincere apology plus reparative acts can “reopen” social channels. Corporate compliance systems, academic retractions, and truth-and-reconciliation commissions all institutionalize counterweight pathways that prevent communities from devolving into uncompensable feuds or permanent ostracism.

Prediction (C8): Organizations that maintain formal amends protocols—clear steps to acknowledge wrongdoing, repair damage, and reintegrate offenders—will show shorter conflict cycles, higher member retention, and lower burnout rates. These effects should be mediated by the *re-expansion of options* for participants: after a conflict and its

resolution, people in these organizations should regain their willingness to collaborate and take on tasks (measurable as increased diversity of task initiation post-conflict). Essentially, a good apology system keeps the group's "game" playable for everyone.

17.3.10 End-of-life rites: last-mile balancing

Many traditions concentrate rituals of confession, absolution, last rites (*vidui* in Judaism, pastoral visits in Christianity, presence of sangha in Buddhism) at the end of life—precisely where LoF predicts the shadow price λ_t to be highest (when H_t approaches 0). These rites orchestrate social repair and self-forgiveness, and ensure practical counterweights (seeking forgiveness, giving blessings or bequests, making reconciliations) can be completed promptly before death.

Prediction (C9): In hospice cohorts, patients who participate in structured end-of-life reconciliation or absolution rituals will show measurable variance compression in their well-being (HCl) in the final days or weeks, compared to those who do not. They should also exhibit increased reconciliatory communication (more phone calls or messages mending relationships) and reduced terminal agitation or despair. These effects should hold even when controlling for analgesia (pain meds) and non-denominational spiritual support, indicating it's the *repair process itself* driving the calm closure.

17.3.11 How to test the cultural echo hypothesis

To rigorously evaluate whether these cultural practices truly echo LoF dynamics, one could:

Corpus linguistics: Quantify moral-accounting metaphors across many languages and historical texts; link the prevalence of such metaphors to societal indicators like conflict persistence and variance in well-being or health.

Ritual "experiments": Use pre/post measurements (HCl, heart-rate variability, cortisol levels, social interactions) around events like confession, atonement ceremonies, or mediations. Include control groups or comparison periods via pre-registration.

Policy evaluation: Where restorative justice programs are adopted (randomly or quasi-randomly by jurisdictions), track outcomes like the decay rate of retaliatory acts in community networks and the tilt of offenders' "menus" toward repair vs. recidivism. Compare to purely punitive systems.

End-of-life protocols: Compare hospice or elder-care units that have strong reconciliation facilitation (structured life-review, family meetings to forgive) versus those with standard care. Track markers of ledger compression (see Chapter 11 for metrics on emotional resolution) and agitation.

17.3.12 What cultural echoes are not

These patterns are not proof of any cosmic design or divine intervention, nor are they excuses to tolerate injustice in hopes of future balance. They are convergent inventions—individually evolved cultural solutions—that keep individual and collective streams away from uncompensable extremes. In doing so, they functionally align with what LoF and QS predict should be helpful. If LoF is correct, we expect such inventions to be widespread, explicitly repair-focused, horizon-sensitive (often heightened at life transitions), and demonstrably efficacious at shrinking emotional variance. If LoF is wrong, these features should not systematically co-occur, or they should fail to produce the expected effects under rigorous measurement.

Takeaway: Humanity’s oldest stories about balances and debts can be re-read in modern control terms as *culture-level implementations of counterweights*. They “work”—not because any culture has stumbled on ultimate metaphysical truth, but because communities that engineer repair keep more options open for more people, especially when time is short. That is the survival logic of compensability, and the secular echo of a deeper law. Indeed, rituals of atonement, jubilee debt cancellations, and even public health campaigns for social connection can be viewed as culturally discovered affordance editors: they re-open compensable paths when collective ledgers are strained. LoF treats these not as moral miracles but as low-cost policy levers that widen $\mathcal{A}_i(t)$ for many individuals at once.

17.3.13 Where we go next:

In Section 17.4, we broaden the scope across species. We move from human cultural narratives to cross-species predictions: we will propose tests in other animals (and across different evolutionary contexts) to see which creatures might show end-of-life calming or other LoF-like signatures, and which might not, allowing empirical separation of mere resilience from a true fairness law.

17.4 Cross-Species Predictions

If the Law of Fairness is genuinely *law-like*, its signatures should not be unique to humans. They must appear—scaled and refracted—across animals wherever subjective experience plausibly occurs. Below are concrete, testable predictions organized by domain, with suggested methods and clear disconfirmers noted for each. These predictions are designed to avoid anthropomorphism and stick to measurable behaviors and physiology.

17.4.1 Behavioral ledgers in the wild

Prediction A: Compensation-biased choice under short horizons. In species with observable horizon changes (e.g., seasonal die-off approaching, terminal illness, or post-injury decline), ethological observations should reveal a shift toward reparative or load-reducing behaviors as horizons shrink. For example: seeking shelter more often, increasing affiliative contact (grooming, huddling), engaging in analgesic self-care (dust-bathing, mild self-stimulation), and conserving energy (less risky foraging, more rest).

Method: Longitudinal focal animal observations with horizon proxies (e.g., body condition score, parasite load, known seasonal timelines). Pre-register which behaviors count as “compensatory” and compare their frequency in periods well before an expected horizon contraction vs. right before it.

Disconfirmers: No shift, or an opposite shift (more risk-taking, less repair) as horizons shrink, *despite* compensatory options being available and low-cost. If animals near death or seasonal endpoint don’t bias toward easier, comfort-enhancing activities, that would undermine the LoF expectation.

Prediction B: Counterweight play. In juveniles of social mammals and intelligent birds (corvids, parrots), play behavior (especially social play) and reconciliation gestures should increase following days with elevated stressor load (e.g., a cold snap, social defeat, predator scare), even if caloric intake is unchanged. The idea is that after a “bad day,” youngsters naturally seek extra counterweights (play being a positive, learning-rich experience, and social play repairing bonds).

Method: Use accelerometers and social network analysis on wild juveniles. Quantify rough-and-tumble play vs. solitary behavior, and relate changes to prior-day stress metrics (temperature, aggression events). Blinded observers can code affiliative play events.

Disconfirmers: Flat or negative coupling between prior-day stress and next-day affiliative play *when* safety and energy availability are adequate. If stress doesn’t lead to more play

or soothing contact in juveniles, LoF's idea of natural counterweight insertion is challenged.

Prediction C: "Menu shrinkage" near closure. As an animal enters end-of-season or end-of-life phases, its behavioral repertoire (diversity of behaviors) should narrow to fewer, higher-Φ acts, and risky high-cost sequences should be aborted more frequently. In other words, the closer to the end, the more the creature sticks to basics (e.g., resting, grooming, feeding) and avoids novel or high-variance behaviors.

Method: Calculate repertoire entropy or diversity over time for individuals. Use survival analysis linking declines in behavioral diversity to independent horizon markers (like known age, or biomarkers of senescence).

Disconfirmer: If repertoires stay the same or even broaden despite shrinking horizons, that would indicate no built-in "last lap" adjustment, contradicting the LoF expectation.

17.4.2 Sleep and dreams as low-cost counterweights

Prediction D: Valence inversion in REM-like states. Species with REM/NREM sleep (mammals, many birds, perhaps reptiles) will exhibit increased REM duration or intensity after aversive days. Moreover, dream-content proxies (e.g., muscle twitches, replay of learned stimuli, neural reactivation patterns) will skew toward reframing stressors rather than amplifying them. For instance, an animal stressed by a new predator scent might show intensive REM and then approach that scent more calmly after sleep (as if nightmares or counter-dreams defused it).

Method: Polysomnography on animals subjected to mild stress vs. control, ensuring ethical bounds. Use closed-loop techniques to interrupt REM at times and see if that affects next-day behavior (e.g., approach to previously avoided cues). Also monitor neural activity during REM for patterns related to waking stressors.

Disconfirmer: No systematic REM change after stress, or evidence that dream replays *amplify* fear (leading to worse avoidance) rather than rebalance it. If stressed animals either don't dream more or dream in a way that *increases* subsequent fear, that contradicts the idea of dreams as QS counterweights.

Prediction E: REM protection under scarcity. Even under caloric restriction or other resource scarcity, animals will conserve a minimum REM quota when affective "debt" is high. In other words, if an animal is both hungry and stressed, it will trade off other activities (like foraging time) to still get enough REM sleep, rather than curtail REM to save energy. This implies REM is treated as a non-negotiable compensation layer.
Method: Use a factorial design: impose mild food restriction and/or stress. See if animals

under stress *maintain* their REM sleep more than non-stressed, even at the cost of less feeding or less NREM.

Disconfirmer: If, under combined stress and scarcity, REM is the first thing the animal sacrifices (with no alternative mechanism stepping in), then REM isn't acting as the crucial counterweight LoF suggests.

17.4.3 Neural homologies of admissible-set control

Prediction F: Control-hub modulation across clades. In mammals, regions like ACC (anterior cingulate cortex) and rIFG (right inferior frontal gyrus) should show stronger inhibitory gating on low- Φ options as horizons shrink (e.g., an animal nearing winter or infirmity shows more ACC engagement to suppress risky moves). In birds, an analogous pattern should appear in their nidopallium caudolaterale (a higher cognition center); in cephalopods, circuits in the vertical lobe (learning/memory center) might show a compensatory bias via neuromodulator shifts. The expectation is that nature uses different hardware to implement the same functional constraint.

Method: fMRI or fiber photometry in rodents performing tasks with horizon manipulation; immediate early gene (c-fos) mapping in birds given short vs. long future contexts; calcium imaging in octopuses during choices with different risk/reward timing.

Disconfirmer: If neural signals in these control hubs track only reward or general arousal, with no independent $\Phi \times H$ interaction (no special signal for “this choice endangers long-term balance”), then QS-specific modulation is absent.

Prediction G: Neuromodulator “shadow price.” Neurochemicals like norepinephrine (NE) and serotonin (5-HT) should track the theoretical shadow price λ_t of well-being. Specifically, as horizons shorten or ledgers drift negative, we expect rising NE/5-HT tone that biases the brain toward caution, withdrawal, and repair. This would be a phylogenetically old way of implementing “watch out, you’re in debt” signals.

Method: Use microdialysis or biosensors in animals during tasks where horizon and ledger are manipulated (e.g., give an animal a series of losses and a time limit). Also, pharmacologically clamp NE or 5-HT levels to see if it disrupts normal horizon-related adjustments.

Disconfirmer: If NE and 5-HT simply track generic dimensions (like surprise or reward) and show no extra kick when an animal is low on time and failing, then they aren’t carrying a shadow price signal. A purely monotonic coupling to stress or reward, without the predicted *horizon × debt* dynamic, would count against the theory.

17.4.4 Developmental scale and lifespan.

Prediction H: Faster lives, faster balancing. Short-lived species (e.g., zebrafish, fruit flies if they have rudiments of affect) should show steeper horizon scaling—i.e. they shift to compensatory behaviors much more quickly and perhaps more extremely as they age—because they have to finish balancing within a short lifespan. For example, a fruit fly (if conscious at all) might have an intense “retirement” phase compressed into a few days. *Method:* Compare compensatory behavior slopes across species with varying lifespans, controlling for metabolic rate. For instance, measure how quickly older animals shift to low-risk, high-comfort behaviors after reaching a certain age fraction.

Disconfirmer: If a fly and an elephant have identical patterns of behavior change with age (when scaled by percent of lifespan) despite radically different lifespans, then LoF scaling isn’t showing up. We expect differences: short horizon species should hit the compensatory mode sooner and harder.

Prediction I: Sensitive periods as balance points. After developmental insults (e.g., a period of scarcity or trauma early in life), juveniles might show an overshoot counterweight shortly after that window if safe channels are present. In other words, a young animal might engage in extra play or social bonding as soon as conditions improve, as if to “make up” for the bad time. This would reflect an internal drive to correct a ledger that went too negative during a critical period.

Method: Ethically implement a mild, time-bounded stress during a known sensitive period, then provide an enriched recovery environment. Measure the magnitude and latency of affiliative or joyful behaviors once the stressor is lifted.

Disconfirmer: If, upon relief, juveniles do not rebound with any extra positive behaviors (they just proceed normally), then the idea of compensatory overshoot for missed balance is not supported.

17.4.5 Social species and network-level repair.

Prediction J: Ledger-aware grooming. In highly social species (primates, social carnivores, parrots), grooming or food-sharing should bias toward individuals who have recently suffered deficits (e.g., lost a fight, got injured), *provided* helping them is low risk (no contagion, etc.). Essentially, social animals unconsciously “redistribute” comfort to those running a hedonic deficit, which benefits group stability.

Method: Analyze social exchange networks with a memory of recent negative events. Use hidden Markov models or Bayesian models to see if an individual’s probability of

receiving grooming goes up after they've had a bad day. Control for rank and kinship to isolate the effect of need.

Disconfirmer: If animals consistently ignore or even attack high-deficit individuals when aid would cost them little, then there's no sign of an evolved bias to compensate others' ledgers (beyond simple reciprocity or alliance considerations). LoF wouldn't be manifesting at the group level.

Prediction K: Coalition restraint near closure. In group-living species with dominance hierarchies, dominants approaching the end of a season or end of life will initiate fewer aggressive escalations and more reconciliations, compared to their mid-life behavior. The idea is that as a dominant's horizon shrinks, they "soften," prioritizing peace over continued fights, since they can't use future time to make up for new injuries or grudges.

Method: Track identified dominants across seasons or across their lifespan, noting rates of conflict they start vs. reconciliation behaviors they perform. Compare late-stage vs. prime-stage frequencies, controlling for physical condition.

Disconfirmer: If older dominants remain just as aggressive or more so, with no uptick in conciliatory acts, then the horizon effect on social strategy is not present (or is swamped by other factors like senility or desperation).

17.4.6 Predator-prey and harsh ecologies

Prediction L: Event-coupled counterweights. Predators that experience a failed hunt (a negative event) will show increased social contact or rest afterward that goes beyond simple energy conservation predictions. Similarly, prey animals that survive a near-miss chase will engage in extra bonding or soothing activities once safe (like group resting or allogrooming). The logic: a brush with adversity triggers a quick counterweight—"take it easy, recover, and bond" before resuming normal activity.

Method: Use bio-loggers for heart-rate and activity on wild predators/prey. Identify high-stress events (spikes in HR, accelerometer bursts), then examine the subsequent window for unusual levels of rest or social behavior. Compare to days without such events.

Disconfirmer: If after a scare or failure animals immediately go back to business-as-usual with no extra rest or contact (beyond what an energy model would dictate), then this predicted counterweight behavior is not robust.

Prediction M: Catastrophe corridors. After major environmental shocks (wildfires, droughts, etc.), if minimal safety and nutrition are restored, animals across taxa should show a rapid rebound in reparative behaviors—not just feeding but social soothing,

increased affiliative calls, and cooperative care. This is a population-level version of the counterweight idea: after a collective trauma, a burst of prosociality and restabilization attempts.

Method: Observe wildlife in the aftermath of natural disasters or sudden resource bonanzas following a dearth. Compare sites or populations that experienced shock vs. those that didn't, for behaviors like grooming rates, vocal reunion calls, sharing of resources. Cross-site replication strengthens the case.

Disconfirmers: If communities remain in a suppressed state of repair behaviors long after the external stressor is removed (i.e., they don't rebound socially when they could), then the hypothesized compensatory drive isn't kicking in.

17.4.7 Edge taxa and distributed minds

Prediction N: Cephalopod compensation without cortex. Octopuses (and some other cephalopods) have high problem-solving abilities but very different brains (distributed nervous systems). If they have subjective experience, we should still see counterweight behaviors: e.g., after aversive handling, an octopus might initially avoid the handler but later show exploratory approach (as if curiosity returns), along with restorative stillness periods that aren't simply fatigue. Also, the vertical lobe in their brain (learning center) may show plasticity correlates indicating adaptation to offset the stress.
Method: Handling stress experiments with careful welfare protocols. Measure changes in exploration of a previously threatening environment and use calcium imaging or electrodes to see if vertical lobe activity changes after stress vs. after rest.

Disconfirmers: If octopuses (assuming they feel something) either never explore after stress (just remain fearful until death) or don't have any quiet "recovery" behaviors that seem beneficial beyond rest, then maybe their control system doesn't do what QS would predict. Also, if any observed compensation can be fully explained by fatigue (with no active re-engagement later), it's not a true counterweight.

Prediction O: Insect minima (a sentience test). In insects with debated sentience (honeybees, some flies), if they do have a minimal subjective experience, they should show short-horizon compensations. For example, after a stress (like mild shaking or CO₂ exposure), a bee might later show a rebound in sucrose responsiveness (a positive bump) or increased social aggregation once safe. If no such pattern is seen, it suggests minimal or no felt experience to regulate. Thus LoF signatures (or lack thereof) could actually serve as a *diagnostic for sentience*.

Method: Stress bees in a mild way and measure changes in sucrose preference, social huddling, or sleep patterns. Compare with truly non-sentient controls (e.g., simple robots or very simple neural organisms) to see if any unique pattern emerges.

Disconfirmer: If robust LoF-like patterns appeared in clearly non-sentient systems, that would be embarrassing (false positive). Conversely, if a species considered likely to feel pain (like octopuses or certain social insects) shows no compensatory adjustments given every opportunity, that's evidence either against their sentience or against LoF's generality.

17.4.8 Cross-species HCI: how to measure without anthropomorphism

Prediction P: Convergent composite indices. We should be able to construct a species-tailored Hedonic Composite Index (HCI) for various animals (combining autonomic signals like heart-rate variability, pupil dilation, with behavioral micro-expressions, sleep patterns, task performance changes) and find that these indices reveal ledger drifts and compensatory tilts that can be meaningfully compared across taxa. After normalizing for metabolism and lifespan, the patterns should converge (e.g., all mammals might keep their HCI within certain bounds, just on different timescales).

Method: Develop per-species HCI models using modern machine learning on multimodal data. Then create a common space by z-scoring and scaling by lifespan or metabolic rate. Use preregistered similarity criteria to test whether distributional features are comparable across species.

Disconfirmer: If each species' affect dynamics are idiosyncratic and can't be aligned by any sensible normalization, or if attempts to do so collapse under cross-validation (no general latent structure), then the idea of a universal law influencing them is weakened.

Prediction Q: Uncertainty handling respects LoF. Decision-making under uncertainty should still obey the LoF logic. Specifically, when there is uncertainty in an agent's estimate of its ledger or prospects, its policies will reflect the posterior probability of neutrality, not just expected utility. In practice, that might look like risk-sensitive compensation: if an animal *isn't* sure how close it is to "doom," it will behave conservatively as if to ensure fairness.

Method: Construct Bayesian state-space models of animal behavior in situations with ambiguous cues. Test if adding a "neutrality probability" factor (the probability that current trajectory will allow neutral closure) improves predictions of behavior beyond a standard reward-risk model.

Disconfirmer: If actions track only expected reward and classical risk measures, and adding a neutrality-probability term adds nothing, then animals aren't acting as if they have an internal model of LoF uncertainty. That would suggest no such constraint.

Finally, what success looks like is a coherent multi-species picture: we find (1) clear horizon-contingent compensatory tilts in behavior, (2) sleep-based counterweights and REM effects as predicted, (3) neural and neuromodulatory signatures that go beyond simple utility or stress models, and (4) convergent well-being dynamics across very different animals once properly scaled. What failure looks like is the opposite: animals nearing death who carry on as usual (no special changes), no REM protection after adversity, purely reward/utility-driven neural patterns, and completely species-specific idiosyncrasies with no rhyme or reason connecting them.

These predictions keep LoF non-teleological and strictly testable: they ask not for miracles or moral judgment, but for lawful constraint signatures that should echo wherever experience—and thus fairness—can meaningfully apply.

17.4.9 Where we go next:

Next, Section 17.5 turns to the practical matter of experiment design. Having outlined predictions in the wild, we lay out how to rigorously test LoF in controlled studies. We will describe “fitness-neutral, constraint-binding” experiments that hold survival fitness constant while manipulating emotional balance, to isolate whether a fairness constraint is at work beyond basic viability.

17.5 Research Notes: Fitness-Neutral, Constraint-Binding

Purpose: To clarify how the Law of Fairness (LoF) can be fitness-neutral (it does not systematically raise or lower Darwinian fitness across a lineage) while remaining constraint-binding (it restricts which trajectories a conscious organism can stably realize). The aim is to lay out formal handles, comparative tests, and clear disconfirmers for the evolutionary theory.

17.5.1 Conceptual frame

By fitness-neutral, we mean that averaged over ecological contexts and genotypes, LoF does not confer a directional selection advantage. You can evolve clever strategies or clumsy ones under LoF; the law doesn't help you "win" the genetic game overall. By constraint-binding, we mean that whatever strategies *do* evolve must respect a feasibility cone in affective space. In particular, life trajectories must remain compensable (i.e. admit some path to hedonic neutral closure) under organism-typical horizons and channels.

Analogy: Gravity is fitness-neutral for birds (some birds fly well, some poorly), but it is constraint-binding: *no* bird evolves wings that ignore gravity. LoF should function similarly for streams of experience. It doesn't make organisms happier or more fit on average; it just forbids life trajectories that would break the balance.

17.5.2 Research notes: the "forbidden experiment" (maze-rat)

A scientific hypothesis lives or dies by its ability to make testable predictions. We have claimed something specific and extreme: that every conscious life ends with a balanced felt ledger. So we must ask the hard question directly. What would the world look like if this were not true? What evidence would flat-out falsify the Law of Fairness? Naming these failure conditions is not a formality. It is the discipline that keeps the theory honest. If the law is false, we want to know it clearly.

The most direct way the Law of Fairness could fail is simple in principle: identify a single conscious life that ends with a clear surplus of joy over suffering, or a clear surplus of suffering over joy, beyond any reasonable margin of uncertainty. A truly unbalanced ending, even once, is enough. The law does not get to be true "on average." It must hold for every conscious stream or it is not a law.

What would such a counterexample look like in practice? Consider a child born into extreme suffering who dies young after a short life of agony. If that child experienced nothing that could plausibly counterbalance the suffering before the Death of Mind, the ledger would appear to end in a deficit. On the other side, imagine an individual who lives

a consistently blissful life and dies peacefully without any comparable inner turmoil or counterweight. If their experience remained deeply positive through the end, the ledger would appear to close with a surplus. Either outcome would refute the claim that terminal balance is guaranteed.

At this point it is worth considering how far one might try to push testing in theory. There is one class of experiment that would test the Law of Fairness more directly than almost any other, and for that very reason it is an experiment we must not perform. In principle, with enough resources, one could attempt to engineer a human life that is as close as possible to a pure surplus on one side of the emotional ledger: either a life of sustained suffering with every possible source of comfort, relief, or meaning intentionally removed, or the reverse, a life of continuous pleasure, protection, and fulfillment with every significant hardship systematically excluded. If such a life could be created, carefully monitored, and followed all the way to the end of consciousness, the Law of Fairness would face an unusually sharp test. A clear surplus of pain or pleasure at life's end would falsify the law outright. Conceptually, it is powerful. Morally, it is indefensible. Trapping someone in an experiment from birth to death and intentionally constructing extreme imbalance, especially suffering, would be a profound violation of ethical norms and human dignity. A theory of fairness cannot justify cruelty in the name of its own verification.

A deliberately extreme version of this logic appears strictly as a thought experiment. In this scenario, three groups of mice are defined: a baseline group, a group preloaded with sustained positive experience, and a group preloaded with sustained negative experience. Each subject is then placed into a one-time maze containing one hundred exits. Only one exit preserves life. The remaining ninety-nine result in immediate death, terminating all future experience and eliminating any possibility of later compensation. Under pure chance, the probability of survival is one percent.

The point is not to advocate such a study. It is to stress-test falsifiability under a hard horizon. If only one path preserves any future at all, then any mechanism that forbids unresolved imbalance would, in theory, have to bias outcomes toward that path, especially for streams whose emotional ledgers could not otherwise close lawfully. To sharpen this stress test, the scenario is intentionally strengthened with adversarial features: large cohorts chosen randomly that are never replenished, strict time limits that prevent gradual learning, misleading cues on fatal exits, and no lure or reinforcement at the safe exit. Precisely because it collapses fairness, survival, and finality into a single irreversible choice, it represents a conceptual extreme. It is also ethically unacceptable outside the realm of thought.

For that reason, “forbidden experiments” are not proposals. They are boundary markers. They clarify what decisive disproof would look like, while also reminding us that not all imaginable tests are permissible. The challenge is to find ethical analogues: natural experiments, observational studies, end-of-life research, longitudinal tracking, and carefully designed behavioral or computational models that probe imbalance and compensation without manufacturing harm.

A different stress test challenges the Law of Fairness from the opposite direction. Instead of trying to force a terminal deficit of suffering, it asks whether a conscious life can be sustained in a persistent surplus of positive experience without triggering compensatory correction.

Imagine, as a theoretical construction, a cohort of lives placed into environments designed to minimize hardship rather than impose it. These environments are not chaotic or indulgent, but stable, calm, and protective. Conflict is minimized. Basic needs are reliably met. Social interactions are structured to be supportive rather than adversarial. Effort is reduced but not eliminated. Meaning is provided without prolonged struggle. From the outside, these lives would appear unusually fortunate.

The critical question is not whether such lives would feel pleasant in the short term. The question is whether they could remain experientially unbalanced over time.

If the Law of Fairness is false, nothing prevents a life from accumulating a lasting surplus of pleasure under these conditions. The absence of deprivation would simply produce a happier trajectory. No correction would be required. No internal resistance would need to emerge.

If the Law of Fairness is real, the prediction is starkly different. Under the law, sustained insulation from hardship cannot preserve experiential imbalance indefinitely. Even in the absence of external adversity, compensatory mechanisms must arise from within the conscious stream itself—particularly in those lives that naturally terminate during the observation period.

Crucially, the law predicts a correlation between terminal outcomes and internal correction. Lives that conclude while externally protected should nevertheless exhibit measurable compensatory dynamics prior to closure. These may take subtle but detectable forms: emotional flattening, loss of salience, diffuse dissatisfaction, interpersonal friction, boredom, anxiety without an identifiable cause, or heightened sensitivity to minor disturbances. The surplus would not simply disappear; it would become unstable, eroding from within until balance is restored or the stream terminates.

Crucially, the Law does not predict punishment. It predicts correction. If the Law of Fairness exists, it does not require cruelty, trauma, or imposed suffering to restore balance. It requires only that no conscious stream be permitted to terminate with unresolved excess. If external hardship is suppressed, internal processes must carry the load.

This thought experiment therefore places the Law under a complementary constraint. The forbidden experiment asks whether a life can be ended before fairness is achieved. This engineered surplus case asks whether fairness can be indefinitely postponed by protection alone. If carefully insulated lives develop compensatory negative experience despite the systematic absence of overt adversity, then fairness is not merely an artifact of environment or upbringing. It is a structural property of conscious experience. If, on the other hand, such lives remain stably positive across time without internal correction and then terminate in that state, the Law fails. A true surplus has been demonstrated.

This experiment is not a proposal. It is a boundary condition. It defines what would have to be possible for the Law of Fairness to be false, and what must be impossible if it is true.

Beyond these boundary tests, we could also disprove or weaken the law by looking at patterns across many lives, especially in extreme circumstances. The law asserts universality. So, if we found an entire category of lives that systematically end positive or negative, that would be evidence against it. If, even after accounting for plausible psychological adaptation and measurement uncertainty, a stable imbalance correlated with factors like extreme oppression, severe chronic illness, or enduring privilege persisted at the end of life, the premise of an invariant law would be false.

Another test is to examine the end-of-life period in detail. The theory predicts a convergence toward neutrality as death approaches. Concretely, we would generally expect increased efforts to resolve emotional debts: reconciling, seeking forgiveness, making amends, letting go of long-held grievances, narrowing attention toward what matters most. We might also expect affect to move away from extremes and toward a more tempered state. If, instead, careful observation repeatedly found lucid individuals dying with large unresolved emotional imbalance, with time to spare and no sign of convergence, that would be evidence against the law.

We can also test whether the law adds explanatory power beyond known psychology. If every apparent balancing effect can be fully explained by standard mechanisms and random variation, and we still find clear end-of-life outliers, then the Law of Fairness loses its necessity. If there is no detectable convergence toward neutrality, and if some lives clearly end far from balance, then the law fails.

The conceptual setup (thought experiment only).

- Groups: Three groups (G1 normal baseline; G2 preloaded with sustained positive experience; G3 preloaded with sustained negative experience). Preload refers to documented differences in affective state, not to deserts or moral judgments.
- Test arena: A one-time terminal maze with 100 exits, one exit safe (continued normal life), 99 exits terminal (no possibility of later compensation).
- Procedure: A rat is placed at the center. If it reaches the safe exit, it lives; otherwise, it does not.
- Why this would be decisive if it were permissible: Under chance alone, safe-exit probability is 1%. Under LoF as a constraint on admissible trajectories, if a stream is severely imbalanced and the environment suddenly imposes a hard horizon (no future steps if the wrong exit is chosen), LoF predicts that admissible policies would narrow toward those that preserve future compensation feasibility. In this contrived endgame, only the safe exit preserves the possibility space for later neutrality; all others foreclose it. Under LoF the probability mass assigned to the safe exit should increase relative to chance for streams whose ledgers cannot otherwise lawfully close. That is the prediction the thought experiment is meant to sharpen—not to enact.

To strengthen the theoretical experiment:

- Pre-commitment and cohort tracking. Each group contains 1,000 rats in identical preloads (G1 “normal,” G2 positive, G3 negative). Before birth, 100 in each cohort are randomly pre-committed as “entrants,” and all 1,000 per cohort are tracked from birth (behavior + brain state sampling). Entrants are never replaced. If any entrant dies before the maze day, we do not substitute a reserve.
Why this matters conceptually: Under a strict balance constraint, streams whose ledgers could not be lawfully closed under the maze’s hard horizon should become less likely to proceed into the terminal episode or will exhibit materially different state trajectories relative to non-entrances within the same group. In this contrived story, improbable “pre-maze attrition” among pre-committed entrants would be read as a narrowing of admissible trajectories before the endgame.
- A hard time horizon. If an entrant does not choose an exit within 10 minutes, the episode ends terminally (no continuation). *Why this matters conceptually:* It turns the horizon into a true “hard wall,” making the Queue System’s horizon-weighting

most stringent—only the unique future-preserving exit remains admissible for streams that cannot close within the neutrality constraint otherwise.

- Adversarial lures. To stress the mechanism, some wrong doors are baited with strong proximal motivators (food scent, social cues), while the single safe door has no lure. *Why this matters conceptually:* If the admissibility filter is real, the policy mass for streams requiring continuation should still concentrate on the single future-preserving exit as the horizon hardens, despite misleading near-term rewards.

Why the real-world version is unethical and scientifically brittle.

- It deliberately harms sentient creatures; that alone settles the matter.
- It collapses *viability* and *fairness* into the same event (death or not), making interpretation circular.
- It is riddled with confounds (scent trails, social learning, stress responses, handling differences, path-length asymmetries) that could swamp any subtle QS effect.
- Most importantly, it removes the very future that LoF reasons about. A one-step “maze of death” is not a fair representation of ordinary admissible life trajectories.

What we test instead (ethical, decision-grade proxies). We preserve the logic (compensation must remain feasible; hardening horizons intensify constraint) while eliminating harm and circularity. Large-N non-sentient agents (primary proxy).

- World: A procedurally generated “labyrinth” with 100 exits per episode; 99 exits reset the agent and impose a cost in the internal objective; 1 exit preserves an open future (continued episodes with cumulative reward). No sentience is involved; these are software agents.
- Preload: Before the terminal episode, agents undergo long pre-training that yields three distributions over their cumulative internal return (analogs of G1/G2/G3). We hold survival/episode length constant across groups (“fitness-neutral”), changing only the affective objective analogue.
- Prediction under LoF-style constraints: In the terminal episode, agents with strongly imbalanced internal return histories should show a higher policy mass on the unique exit that preserves future opportunities.
- Model comparison: Compare a LoF-constrained policy family (QS-style admissibility filter with horizon-weighted cost term λ_t) to standard RL baselines

that optimize myopic or fixed-horizon returns. Name LOO (leave-one-out) once here as the chapter’s out-of-sample metric for predictive performance and falsification. [OOS metric already named earlier in this chapter]

- Fail pattern (strengthened). If, under hardened horizons and lures, constrained agents do not increase policy mass on the continuation exit relative to tuned baselines (after hyperparameter sweeps), the constraint hypothesis weakens.
- Pre-commitment analogue. At pre-registration, randomly tag 100 candidate agents out of 1,000 per preload condition (analogs of G1/G2/G3). Never replace tagged agents that “attrit” during pre-training. Track whether tagged agents with extreme return histories disproportionately fail to reach the terminal episode under hardened horizons (evidence for admissibility narrowing happens *before* the endgame).
- Hard horizon. Set a strict episode-limit; 99 exits reset the agent and impose a cumulative-return cost; one exit continues the series (open future).
- Adversarial lures. Place shaped local rewards at reset-exits; leave the continue-exit unrewarded locally. Prediction: LoF-constrained policies shift mass to the continuation exit as the horizon tightens, despite local lures.

Human, low-stakes virtual maze (behavioral proxy).

- World: A browser task with 100 “doors” per round; 99 end the round with a time-loss and a small charitable donation not made; 1 door continues the session and allows the donation to go through. No deception about harm or personal loss.
- Preload: Participants come in after preregistered periods of unusually good or bad days (documented with HCI sampling) or randomized vignette preloads that move HCI briefly but ethically.
- Prediction: Under a short-horizon frame (few rounds left), streams with recent negative load should more often select the future-preserving option (greater willingness to work for continuation), relative to matched controls.
- Pre-commitment analogue. Pre-tag a subset of participants before the session (randomly selected and preregistered), and analyze them as a fixed cohort even if some withdraw; compare to matched controls to guard against attrition bias.
- Hard horizon. Use brief rounds with “few moves left” timers; failure to choose ends the round with a small charitable donation not made.

- Adversarial lures. Place time-saving or tempting cues on losing doors. Prediction: Under short horizons, recent negative-load streams show higher rates of choosing the future-preserving option despite lures, vs. matched controls.
- Measurement: Choice counts follow Poisson; if dispersion is meaningfully > 1 , pivot to Negative Binomial and state the link. Report configural \rightarrow metric invariance checks for any cross-group claims; if metric fails, restrict to within-person.

Physical robots, zero harm (engineering proxy).

- Simple mobile robots in a maze with 1 “continue” exit and 99 “reset to dock” exits. “Affect” is an internal energy-opportunity budget, not suffering. Under shortened horizons (battery nearly depleted), LoF-constrained control should increase routes that reach the continue-exit vs. standard controllers.
- Pre-commitment analogue. Tag a subset of robots as “entrants” at build time; do not swap units if batteries or components fail prior to the terminal run; analyze whether constrained controllers admit fewer “non-compensable” trajectories into the last run.
- Hard horizon + lures. Use a battery-limited run with 99 “reset to dock” exits (local energy/latency rewards) and one “continue” exit. Prediction: Constrained control increases routes that reach “continue” despite local lure rewards.

What would count against LoF in this family.

- No horizon effect: tightening horizons fails to change policies in any proxy.
- No QS signature: after utility, difficulty, and prediction-error are modeled, there is no residual tied to compensability Φ or to λ_t in constrained agents.
- Expanding rather than compressing variance near termination windows across agents or participants.

Any of these—documented with preregistered bounds—counts as evidence against the constraint.

Why the thought experiment still belongs in the book. It forces precision. It shows *exactly* where a law-level claim must cash out: when all but one path foreclose future compensation, admissible trajectories, if the law is real, must concentrate on the single path that keeps neutrality possible. We then show how to test that structure ethically and rigorously.

Math note (notation for preregistration).

Inline ledger and estimate:

$$L(T) = \int_0^T F(t) dt$$

$$\bar{L}(t) = \int_0^t HCl(\tau) d\tau$$

Constraint-weighted admissibility (sketch): policies in $\mathcal{A}(t; \bar{L}, H, C)$ must preserve the feasibility of neutrality as T approaches; the horizon gain $\beta(H_t)$ increases as H_t shrinks.

17.5.4 A minimal formalization (replicator + constraint)

Consider a population with strategies indexed by i . Let $x_i(t)$ be the frequency of strategy i at time t , and let f_i be its Malthusian fitness (growth rate), with \bar{f} the population mean fitness. Standard replicator dynamics are: $\dot{x}_i = x_i (f_i - \bar{f})$.

Now introduce a fairness constraint. Let each strategy induce, for typical lifetimes, a neutrality probability $P_i \equiv \Pr(|L(T)| \in [-K, K] | \text{strategy } i)$, i.e., the probability that a life governed by strategy i ends with a ledger within $[-K, K]$. In a strict per-strategy formulation, LoF imposes a hard admissibility requirement $P_i \geq 1 - \varepsilon$ (for some small ε). Then the feasible set of strategies is:

$$\mathcal{F} = \{ i : P_i \geq 1 - \varepsilon \}.$$

Population dynamics effectively unfold on the restricted simplex $\Delta_{\mathcal{F}}$ (only the allowed strategies) rather than the full strategy simplex.

Alternatively, in a relaxed population-level formulation, one can view this as an optimization with a constraint: maximize mean fitness subject to average fairness feasibility. In a simple form:

$$\max \{x_i\} \sum_i x_i f_i \text{ s.t. } \sum_i x_i P_i \geq 1 - \varepsilon, \sum_i x_i = 1, x_i \geq 0.$$

The Kuhn–Tucker multiplier λ^* on the fairness constraint acts as a population-level shadow price of fairness (cf. the individual-level λ_t in Chapters 5–6). Importantly: Fitness-neutrality test: At equilibrium, selection gradients along any fairness-preserving direction are unchanged; LoF only truncates illegal directions. Formally, for any feasible perturbation vector v such that $\nabla P \cdot v = 0$ (no change in fairness outcome), the fitness gradient $\nabla f \cdot v$ is unaffected by the constraint. In other words, within the allowed cone, evolution proceeds as usual.

Constraint-binding test: There exist directions w in strategy space for which $\nabla P \cdot w < 0$ (moving that way would reduce P_i). Those trajectories are removed from evolutionary play regardless of their raw fitness payoff. This is the key: some strategies could in principle out-reproduce others, but they never get to, because they violate the fairness bound.

Disconfirmer: If empirical strategy distributions in the wild require violations $P_i \ll 1 - \varepsilon$ to explain observed success (i.e., the only way we can model some species' evolution is to allow that many individuals end life with big uncompensated debts), then LoF is not binding in that context. Essentially, if we find stable lineages where individuals routinely die miserable (or euphoric) with no balancing, then the law as stated doesn't hold.

One well-documented conscious stream that terminates outside preregistered neutrality bounds, after all measurement corrections and channel logging, is sufficient to falsify the Law. The claim is universal or it is not a law.

17.5.3 What “fitness-neutral” would look like in data

How would we know if LoF truly doesn't budge overall fitness?

No consistent selection on ledger offset. Across cohorts and environments, an individual's baseline ledger drift (their average hedonic offset, see HCl in Chapter 7) should *not* predict their reproductive success once you control for obvious confounds (resource access, health, status). Nature isn't systematically favoring the chronically happy or sad *if* LoF holds.

Design: Multi-year studies in social mammals or birds, combining telemetry of behavior/physiology with pedigree data. Use mixed-effects survival or reproductive models with each animal's ledger drift as a covariate, plus controls.

Disconfirmer: A clear monotonic relationship: e.g., individuals with chronically high positive affect have more offspring (or vice versa) even after controlling for other factors. If being an optimist (or pessimist) consistently yields more babies, LoF might not be neutral (it would mean selection is pushing hedonic tone one way).

Equal evolvability within the feasibility cone. Traits that shift *how* organisms achieve compensability (like sleep architecture, affiliative repair behaviors, cognitive reappraisal ability) should show normal genetic variation and respond to selection pressures; whereas traits that would push outside the feasibility cone (say, a mutation that causes total loss of REM sleep in a species that needs it, or a behavior that blocks all social comfort) should either be strongly canalized (developmentally buffered) or lethal. In other words, evolution can play freely *inside* the lines but hits brick walls at the edges.

Design: Use quantitative genetics on traits like amount of REM rebound or tendency to seek friends when stressed, and attempt artificial selection. Also examine natural mutation databases for evidence of conservation (GERP/PhastCons) on genes tied to compensation (oxytocin pathways, etc.).

Disconfirmer: If it's trivially easy to breed a line of animals that completely abolishes a compensatory channel with zero viability cost (e.g., a line of rats with no REM sleep that live fine and reproduce fine), then those channels are not actually binding constraints. LoF would struggle to explain why evolution hadn't already dropped "costly" buffers like that.

Environment-dependent neutrality. LoF doesn't promise everyone is always neutral—just that the *capacity* for neutrality is maintained given available channels. In harsh ecologies vs. benign ones, the mix of compensatory mechanisms under selection may differ (e.g., desert animals rely more on behavioral withdrawal, rainforest animals on social support). But in both contexts, the *aggregate neutrality rate* (the fraction of individuals that finish life near hedonic zero) should stay within a narrow band once safe channels exist. Put another way, LoF predicts that when environments permit, populations will achieve roughly similar fairness outcomes, even if through different means.

Design: Cross-site comparisons—e.g., the same species in a tough environment vs. an easier one. Give them analogous "hospice" or recovery opportunities (like provisioning or safe refuges) and measure end-of-life well-being distributions.

Disconfirmer: Large, directionally stable differences in terminal affect between populations that cannot be erased by providing missing channels. If one site's individuals die consistently happier or sadder even after ensuring they had opportunities for rest, sociality, etc., then something other than LoF is driving that.

17.5.4 Comparative physiology: conserved "compensation stack"

LoF implies evolution will often arrive at a tiered compensation architecture that appears independently in many clades. We might predict layers like:

Layer 1 (fast autonomic): quick rebounds in heart-rate variability (HRV), engagement of the parasympathetic "vagal brake," and balanced norepinephrine/serotonin release to counter shocks.

Layer 2 (sleep-based): adjustments in sleep after stress—e.g., REM increases or flips valence of dreams, and NREM slow-wave activity intensifies after heavy strain (to restore equilibrium).

Layer 3 (social repair): behaviors like grooming, reconciliation, food sharing, or vocal soothing that kick in after social or physical stressors.

Layer 4 (cognitive reframe): in large-brained species, deliberate reappraisal, finding meaning in adversity, or storytelling as a way to cope (a purely mental counterbalance).

Comparative test: Using a phylogenetic comparative method (like PGLS regression), show that distant lineages with credible sentience nevertheless exhibit at least two of these layers. For instance, elephants, dolphins, and ravens might all have some analog of Layers 2 and 3.

Disconfirmer: Discover many sentient-credible taxa that lack *any* recognizable compensation layers despite abundant “easy” opportunities. If, say, certain mammals or birds never sleep more after stress, don’t comfort each other when hurt, and show no internal rebounds—and they thrive anyway—then the supposed necessity of these layers is falsified.

17.5.5 Evolutionary game theory: admissible policy space

We can also frame LoF in game-theoretic terms. Consider a repeated social dilemma (like Hawk–Dove or Prisoner’s Dilemma) where interactions have affective spillovers (the actions affect not just payoffs but also the agents’ hedonic states). Empirically, humans enforce fairness even when it costs them personally. In the classic Ultimatum Game, participants routinely reject highly unequal offers despite losing money (Sanfey, 2003), reflecting the deep inequality aversion formalized in Fehr and Schmidt’s (1999) model of social preferences. This costly punishment of unfairness implies that evolved utility functions incorporate a fairness constraint, paralleling LoF’s prediction that only balance-preserving strategies—those that punish or compensate for large asymmetries—can persist over time. Strategies that maximize expected payoff but generate uncompensable ledgers—for example, a strategy that defects ruthlessly especially late in life, causing huge negative affect to others (and maybe guilt)—would be *inadmissible* under LoF. We can formalize:

Let payoff be $u(a_t, a'_t)$ for actions a_t by the player and a'_t by the co-player at time t , and let the affect update be $F_{t+1} = g(F_t, a_t, a'_t)$ (some function of current affect and actions).

Feasibility under LoF requires that a policy pair (π, π') (strategy for player and co-player) satisfies:

$\Pr(|\sum_t F_t| \leq K) \geq 1 - \varepsilon$, i.e., with high probability, the cumulative affect stays within bounds.

Under this constraint:

Result: The standard set of evolutionarily stable strategies (ESS) shrinks to a subset we can call LoF-ESS. Typically, these will be more cooperative-leaning, repair-capable strategies. Ruthless exploiter strategies that rely on strings of defection—especially near end-of-life when there’s no future to compensate—lose admissibility because when horizons shrink, those strategies crash the ledger. In other words, the presence of a

fairness constraint filters out the “grim” equilibria in favor of ones where agents maintain balance.

Empirical handle: We can test this in human laboratory games by framing some games as having a known horizon vs. indefinite future (see Sections 12.1–12.6 for human data). If LoF applies, we’d predict late-game cooperation upticks or compensatory moves when a neutral closure is salient—beyond what classic reputation or reciprocity models would predict. E.g., players might forgive or give away resources in the final rounds even when it’s not “rational” in standard terms, just to leave with a clean slate.

Disconfirmer: If we see no incremental late-game repair behavior—i.e., people or animals act exactly as traditional game theory predicts without any extra end-of-horizon niceness—then there’s no evidence of an intrinsic fairness dynamic affecting decisions.

17.5.6 Genomic and neurochemical fingerprints (negative as much as positive)

If LoF has shaped evolution, we should find its fingerprints in genomes and neurochemistry:

Canalized minima, not maxima. We expect hard floors in certain traits rather than optimized highs. For example, there might be a minimum necessary REM sleep fraction, a minimum affiliative drive under distress, a minimum analgesic response to pain—below which things go off the rails. But we won’t find special “happiness genes” pushing experiences to maximum; evolution just ensures a baseline of compensation capacity. **Search:** Use constraint-based genomic analyses (GERP scores for evolutionary conserved regions, PhastCons, etc.) around genes responsible for REM maintenance, oxytocin/vasopressin systems (social bonding and repair), interoceptive integration (like insular cortex development). These should show high conservation, implying messing with them is highly disfavored.

Disconfirmer: Find multiple independent lineages of animals that have intact consciousness markers (they likely feel pain, etc.) but have entirely lost these “floors” with no replacement. For instance, an animal that *never* engages in social soothing and seems to do fine, or one that doesn’t need any REM-like sleep. That would show these minima aren’t truly universal constraints.

Neuromodulator guardrails. Neurotransmitter systems (like serotonin, norepinephrine, dopamine) often interact to stabilize mood and behavior. LoF predicts certain conserved coupling motifs: e.g., serotonin levels rising when an individual has a negative ledger *and* a short horizon, acting as a mood stabilizer or behavior inhibitor (this ties back to Prediction G). Likewise, dopamine might be kept in check by stress signals to prevent reckless gambles when one is already down.

Search: Look for cross-species patterns in pharmacology or gene expression. Do very different animals show similar receptor expression patterns linking the “mood” systems to sleep or affiliation circuits? For example, do all mammals have serotonin receptors densely in areas that trigger sleep or social seeking when activated? That would hint at a conserved mechanism to enforce taking a break or seeking comfort when needed.

Disconfirmer: Find dissociations: e.g., a situation where an animal in massive pain and near death actually *lowers* serotonin/NE (instead of raising it) and doesn’t do anything else to compensate. If in some species extreme distress plus limited future doesn’t engage any known neuromodulatory response, then the idea of a built-in guardrail is false for that case.

17.5.7 Parasite manipulation and “cheats” (stress tests)

Some parasites effectively “hijack” host behavior (e.g., Toxoplasma in rodents making them less afraid of cats, Cordyceps fungi in insects causing zombie-like behavior, rabies virus driving aggression). These seem to drive behavior away from self-interest and self-repair. LoF would predict that such manipulations only succeed under certain conditions:

Challenge: Instances where an organism’s behavior is taken over to its detriment (reducing its own compensability) are rare, but illuminating.

LoF prediction: Parasites will succeed primarily when they exploit existing gaps or downturns in the host’s compensation system—e.g., a host with no social support or whose sleep is already disrupted—or when the host’s horizon is anyway short (so maybe it “doesn’t matter” in an evolutionary sense). Otherwise, the host’s countermeasures (like sickness behavior, hiding, sleeping more, avoiding risky areas) should kick in and reassert control. For example, a rabies-infected animal might withdraw if it has a chance, rather than go on a rampage, if conditions allow withdrawal (contrary to the virus’s plan).

Design: Ethically bounded infection studies where we give hosts extra help. E.g., infect a rodent with Toxoplasma but also give it access to a super safe, enriched environment with soothing stimuli and no predator cues. Does it still lose its fear of cats? If LoF holds, maybe the host can resist manipulation better when provided compensatory channels.

Disconfirmer: Cases where a parasite consistently drives its host to self-destructive, ledger-worsening behavior even *when* researchers give the host everything it needs to recover (safe shelter, analgesics, social comfort, plenty of time). If the host still can’t avoid the doom spiral, then the parasite found a loophole in LoF, or LoF isn’t as binding as thought.

17.5.8 Domestication and captive contexts (a natural experiment)

Domesticated animals and zoo animals provide an interesting testbed: they often live easier, safer lives than their wild counterparts, but also sometimes have unnatural constraints (like limited space or social options). LoF predicts that compensation will still occur, but the mix of channels might shift:

For example, a cow or dog might rely more on human-provided social soothing (patting, grooming) and less on foraging-for-pleasure compared to its wild ancestor.

Captive animals might show more of one layer (say, they sleep a lot if nothing else is available) to make up for lack of others (not much social life).

Prediction: Domestication or captivity changes *which* compensation layers dominate, but not the existence of compensation attempts. Domesticated species may lean heavily on social bonding with humans and idle comfort behaviors, whereas wild ones had to use foraging or exploration as relief. Importantly, both should still achieve neutrality on average if channels suffice.

Design: Compare free-living vs. well-cared-for captive conspecifics of the same species. Look at their HCl trends over life, sleep patterns, social grooming rates, etc. Do they end up similarly balanced? Does the captive group perhaps use different “tools” (e.g., more sleep, since food is provided)?

Disconfirmer: Find captive groups with ample channels (good space, social group, enrichment) that nonetheless show chronic drift far from neutral at end-of-life. If zoo animals with all needs met still die significantly depressed or abnormally euphoric, something’s off. (Often, though, issues in captivity are traced to unaddressed needs—this prediction assumes we gave them what they need.)

17.5.9 Simulation sandboxes: evolution inside LoF

We can simulate evolution itself under a fairness law to see what strategies emerge:

Agent-based evolution: Create artificial agents (e.g., reinforcement learning bots) that evolve over generations with an LoF-like constraint. For instance, penalize simulated agents if their cumulative reward (interpreted as affect) goes too high or low given a distribution of lifespans. *Outcome:* We expect a rich diversity of strategies that all stay within the cone. Notably, we predict frequent emergence of counterweight modules—behaviors analogous to rest routines, reconciliation actions, etc., that the agents “discover” as ways to avoid penalty.

Disconfirmer: If optimal populations in the simulation converge on strategies that systematically violate neutrality (perhaps by exploiting some loophole in the model) and do so without consequence, or if they “hack” the constraint somehow and those policies would realistically map to uncompensated lives, then the LoF concept might have theoretical cracks. But if our simulation is fair, that shouldn’t happen.

Ablation studies: In simulation, we can remove a compensation channel and see what happens. LoF predicts that if you remove one channel, *others will compensate* if possible, rather than agents thriving by ignoring compensation entirely. For example, take out “sleep” from the model and maybe agents evolve more social repair behavior to cope, rather than just becoming imbalance-accumulating super-agents.

Disconfirmer: If simulated populations can drop all compensation behaviors and still out-compete those with compensation (contrary to LoF), then *in silico* evolution finds a path to violate neutrality—which suggests nature might too if given enough time.

17.5.10 Clean disconfirmers of “fitness-neutral, constraint-binding”

To wrap up, what evidence would flat-out refute the idea that LoF is just a neutral backdrop constraint? Here are some decisive ones:

Directional selection on imbalance: Find robust, reproducible evidence that *chronic* positive (or negative) ledger drift correlates with higher lifetime reproductive success *after controlling for* everything else. If animals who are consistently happier (or consistently more stressed) have more offspring and nothing balances that, then LoF isn’t neutral—it means nature actually favored an imbalance, violating the neutrality expectation.

Evolvable violation: In a lab, breed a line of a conscious animal (say mice, or fruit flies if we think they feel something) that *lacks* all major compensatory channels and accumulates an uncompensable ledger, yet shows normal viability and no end-of-life balancing. If we can do that via directed breeding or gene editing, and those creatures live out full (but miserable or weirdly blissful) lives and reproduce fine, then LoF is not binding—evolution *could* go there, it just hadn’t yet.

Parasite success under aid: Show that even when researchers provide every conceivable compensatory option (safety, rest, social support, medicine), parasite- or pathology-manipulated hosts *still* engage in self-damaging, ledger-worsening behavior to the bitter end. That would indicate a situation where LoF doesn’t protect the host at all—a direct violation in practice.

Across-taxa absence: Discover multiple credible sentient clades (say, some type of fish, some insects, some mammals) that have no evidence of any compensation layer *despite* abundant opportunities. For instance, find an animal that never seeks rest or comfort even when it could, never adjusts behavior near death, and overall doesn't care about balance—and it doesn't suffer viability issues for it. That would show fairness-respecting designs are not a universal outcome of evolution.

Simulation loophole: In evolutionary simulations or AI training, find that some adversarial evolution scenario produces policy families that violate neutrality while maximizing task reward, and these policies don't trigger any modeled fairness constraint terms—and then those policies, if implemented in real organisms, would presumably also violate LoF. In other words, our best computational search comes up with a loophole to get all the fitness without balancing hedonic books. If that loophole seems biologically plausible and nature could exploit it, LoF would be on shaky ground.

Any one of these would force a serious revision or abandonment of the Law of Fairness as a scientific hypothesis.

17.5.11 What to preregister (nuts and bolts)

As researchers set out to test this theory, some practical points for study design:

Primary outcomes: (i) Neutrality rate at end-of-life (or proxies like final-phase HCl distribution); (ii) Slope of compensatory tilt versus horizon length (how strongly behavior shifts as death nears); (iii) A “REM protection index” after aversive days (does REM sleep rebound); (iv) Social repair bias toward high-deficit individuals (do those who suffered get more care). These should be defined quantitatively.

Controls: Always account for factors like energy balance, disease burden, predator density, dominance rank, etc. LoF effects often masquerade as something else (or vice versa), so controls are crucial to isolate the fairness aspect.

Statistics: Use hierarchical (mixed) models with phylogenetic controls for cross-species data. Employ preregistered ROPE or Bayes-factor thresholds to confirm when something is *not* happening (e.g., that end-of-life affect is statistically indistinguishable from neutral band). For behavioral trajectories, Bayesian state-space models can incorporate uncertainty propagation (since animals don't know precisely when they'll die, etc.).

Open data: Given the complexity, preregistration and open data are vital. Raw ethograms (behavior logs), sleep traces, neuromodulator time-series, and the code for analysis should be shared. This ensures that tests of LoF are transparent and replicable—and any failures to replicate will be visible.

If LoF is truly fitness-neutral yet constraint-binding, then evolution should look creative within limits: nature explores a wide variety of life strategies, but *none* that stably ignore compensability. We should witness conserved guardrails (sleep, soothing, social repair across many species), horizon-sensitive behavioral tilts, and a tendency toward neutral end-states that persists across environments once basic “safety nets” exist. Conversely, systematic, replicable violations of these patterns wouldn’t just tweak the theory—they would falsify its claim to be a binding law of nature.

17.5.12 Where we go next:

We now turn to the decisive test of the evolutionary claim. In Section 17.6, we specify what a genuine species-level violation of the Law of Fairness would look like. Rather than cataloging supportive motifs, we define strict criteria for systematic imbalance and clear disconfirmers. The aim is to separate true violations from ecological artifacts or measurement error. If LoF is constraint-binding, it must survive this audit. If it does not, the law must be revised or relinquished.

17.6 Fail Patterns: Species with Systematic Imbalance

Purpose: To specify what would count as a species-level violation of the Law of Fairness, and how we'd detect it without anthropomorphic bias. We also outline how to separate true failures from ecological or methodological artifacts. A “Fail pattern” here means a reproducible, cross-context tendency for typical members of a species (with credible conscious experience) to end life far from hedonic neutrality *despite* having available, low-cost compensatory channels (sleep, soothing, social repair, analgesia, safety) and non-truncated horizons. In short, a species that seems to systematically break the fairness law.

17.6.1 What “systematic imbalance” must mean (strict definition)

For a species S to exhibit systematic imbalance, all of the following would need to hold true for representative, healthy adults across multiple environments:

Terminal affect drift: The distribution of terminal HCl (or a pre-specified, validated proxy for hedonic state at end of life) is stably offset from zero by a non-trivial amount $|\mu_{\text{terminal}}| \geq \delta$ (where δ is pre-registered and exceeds measurement error), with small variance around that non-zero mean. In other words, most individuals end their life either on the significantly negative side or positive side, not balanced.

Channel adequacy: It can be demonstrated that major compensatory channels were available and utilized (or could have been utilized) at low cost. For instance, the animals had access to safe sleep, could form social bonds, had analgesic remedies (plant-based or endogenous) if needed, etc., and their horizons were not uniformly truncated (i.e., deaths were not predominantly sudden or externally imposed before compensatory processes could operate).

Robustness: The imbalance persists even after controlling for confounds like high disease load, heavy predation pressure, captivity effects, or measurement biases. So it's not just that they all died young from disease (not giving time to balance), or our measuring tool was flawed.

Horizon-insensitivity: The usual horizon-scaling signatures are absent. For example, even as individuals approach the end (or seasonal endpoint), they *don't* show increased repair behavior or narrowing of options—nothing changes to correct the ledger.

All four criteria are required. Anything less can be explained by “ordinary” ecological or physiological factors rather than a violation of LoF.

17.6.2 Candidate taxa and why they tempt false positives

A few types of species superficially seem to violate fairness, but on closer look they may not:

Eusocial insects (ants, bees, termites): These have massive labor asymmetries and entire castes that are “expendable” (like sterile workers who toil until death). It looks terribly unfair (some live only to serve others). However, current evidence for unitary conscious experience in these castes is weak. If workers are more like extensions of a colony “superorganism” than individuals with personal ledgers, LoF might not even apply. *Verdict:* Treat them as non-sentient (for our purposes) controls rather than as tests of LoF.

Cephalopods (octopus, cuttlefish): They have high sentience plausibility (big brains, complex behavior) but extremely short lifespans and often semelparity (one-shot reproduction followed by programmed death). An octopus female caring for eggs will deteriorate and die after the eggs hatch, which looks like tragic imbalance. *Confound:* Their horizons are structurally short due to biology. LoF might be satisfied because a strong late-life tilt does happen, it’s just that life ends shortly after so there’s little time to display a neutral plateau. In other words, they might “tend” to compensate but senescence stops the clock.

Obligate predators in starvation-prone niches (e.g., polar bears): They endure chronic stress and hunger due to extreme environments, which might push ledgers negative on average. *Confound:* Ecological channel scarcity. LoF only promises neutrality *conditional on channels being available*. In a collapsing habitat, all bets are off. So polar bears might show negative end states now because their environment has removed the usual buffers (food, rest in dens, etc.). If you gave them a stable environment or a wildlife sanctuary, they might balance out.

Captive carnivores/primates with stereotypies: Zoo animals sometimes develop repetitive, apparently joyless behaviors (pacing, head-bobbing), suggesting poor well-being. *Confound:* These are often artifacts of channel deprivation—lack of stimulation, lack of proper social groups, disrupted sleep cycles, etc. They’re not a species-level trait; they are fixable with better enrichment. So they don’t invalidate LoF generally, they just show what happens when we block the channels.

Bottom line: Before declaring a “systematic imbalance” species, we must rule out (i) non-sentience of the individuals in question, (ii) inherently short horizons imposed by biology, and (iii) severe channel deprivation artifacts. Only after eliminating those can we consider it a real LoF violation.

17.6.3 Measurement blueprint (how to actually test this)

If we suspect a species might be a true exception, how do we test it rigorously?

Sentience filter: First, ensure the species is likely sentient. Pre-register a conservative criterion for sentience: e.g., demonstrable behavioral flexibility, ability to learn to avoid pain, presence of complex sleep (REM/NREM cycles), and social soothing responses. If a species doesn't clear this bar, LoF may not apply at all.

HCI proxies: Use the Chapter 7 composite (or similar) to measure well-being. This means combining observable behaviors, autonomic measures (heart rate variability, pupil dilation), sleep metrics (like REM rebound), social approach/avoid metrics, and where possible neural indicators (perhaps EEG or functional imaging). We need the best possible read on the ledger over time.

Terminal window sampling: For short-lived creatures, sample intensively over the last 10–20% of their expected lifespan. For longer-lived animals, sample at late-life stages or end-of-season periods aligned with natural mortality risk. Essentially, zoom in on the endgame.

Channel audit: Rigorously document which compensatory channels were available and used. Did the animals have safe places to sleep? Could they engage in social bonding or were they isolated? Could they find any analgesic substances or comforting stimuli? Was there shelter and thermal comfort? We have to show that, objectively, the means to counteract pain or stress were there (or if not, that's an external reason for imbalance).

Horizon framing: Estimate each individual's subjective horizon. This could be by looking at their ecological stability (are things predictable or chaotic?), injury status (a wounded animal knows its horizon might be shorter), or experimentally by providing some risk/survival cues in a lab analog. It's critical to show that not *all* individuals had truncated horizons—some had a “long runway” and still didn't balance, which would be weird.

Analysis: Use appropriate models: likely mixed-effects models with individual and site as random effects if you have multiple populations. Use preregistered interval-based criteria to test whether final HCI falls outside (or remains within) a pre-specified neutral band around zero. And plan for absence-of-evidence issues: define what would count as “no compensatory behavior” in a way that's statistically testable (ROPE again, etc.).

Fail call threshold: We should only call a true LoF failure if multiple independent groups (preferably wild and captive, or different sites) show the same pattern: a directional offset and lack of horizon scaling after a positive channel audit. In other words, if in three

different well-observed situations this species still dies unbalanced, then we have a case. One-off weird observations aren't enough.

17.6.4 What true Fail patterns would look like

If a species genuinely defies LoF, we'd expect to observe things like:

No late-life repair tilt. As mortality risk rises (old age, approaching winter), we do *not* see the usual increases in reconciliation, nest sharing, soothing contact, or comfort-seeking. They keep fighting or struggling or isolating at the end, rather than mellowing out or coming together.

REM non-protection. After aversive days, these animals show no REM rebound; in fact maybe REM might even *decrease* without any other counterweight (like quiet rest) stepping up. Essentially, their physiology doesn't even try the usual tricks to stabilize mood.

Social asymmetry persistence. Let's say in this species some individuals are chronically subordinated or stressed by others. In a LoF world, usually those high-deficit individuals get some compensatory affiliative attention eventually (when the dominants are sated or at peace). Here, maybe not: the lowest-ranking or most harmed individuals continue to get *no* extra grooming or care, even when others in the group have the time/energy to give it. There's no "pity" or evening-out ever.

Ledger variance expansion. Normally, near the end of life, variance in well-being compresses as everyone kind of comes toward a similar neutral zone. In a Fail pattern, we'd see the opposite: as a cohort enters the final stage, the spread of HCl scores actually widens (some are extremely miserable, some maybe oddly euphoric). There's no convergence.

If *all* these hold while we've confirmed channels were available and horizons not abruptly cut, LoF is in serious trouble. That would be a real counterexample: the law fails to account for this species.

17.6.5 How apparent Fail patterns can vanish on closer look

Many times we might think we found a violation, but further scrutiny reveals the species was actually within LoF expectations once hidden factors are accounted for:

Hidden channels. Maybe a species compensates in ways we didn't notice at first. For instance, they might do micro-sleeps or basking in sun (thermal pleasure) or solitary self-soothing behaviors that we misclassified as "idle" or meaningless. Once we measure

those properly, we realize their terminal drift disappears—they were compensating, we just weren’t measuring the right things.

Horizon illusions. Some species have genuinely short horizons built-in (salmon that die after spawning, many insects, some small marsupials). To an observer it may appear they died in imbalance, but there was hardly any time to balance. LoF in such cases predicts a brief intensified counterbalance flurry, not a long neutral period. We might be missing it because it’s quick or subtle. Once we know to look for a short spike of, say, frantic mating or nest-building or something, we might find they did attempt compensation in their little time.

Ecological compression. External factors like droughts, human disturbances, or new predators can “slam channels shut.” If a species is studied under such conditions, they might all look imbalanced. But if you put them in a sanctuary or protected reserve with plenty of resources and no interference, neutrality might return. So what looked like a species fail was actually environment-imposed. LoF says *conditional* on channels, so we have to recreate those conditions to test fairly.

Instrumentation bias. Our HCl or well-being measures might not translate well across species. For example, primate facial expressions won’t apply to octopuses who change color or texture to express states. If we apply the wrong yardstick, we might falsely conclude “this octopus is depressed at the end” when in its own expression it wasn’t. We must calibrate properly (perhaps an octopus turning white-spotted and laying still is actually content, not apathetic). Once the composite index is adjusted for the species, the supposed drift might vanish.

17.6.6 Two worked case studies (hypothetical, instructive)

Let’s run through a pair of imaginary but plausible scenarios to illustrate how to diagnose vs. debunk a Fail pattern:

Case A: Plains baboons under human encroachment. Initial field observations show elderly baboons in a heavily disturbed area end life with negative-skewed HCl: lots of stress behaviors, vigilance, and very little grooming (comfort) in their final weeks. It looks like they die miserable, with few compensatory actions. *Channel audit*: We note this population faces food unpredictability, nocturnal threats (maybe predators or poachers), and grooming time is eaten up by constant predator avoidance. Many compensatory opportunities are absent or costly. *Sanctuary replication*: We take a small group of such baboons to a large sanctuary where food is predictable, safe sleeping platforms are provided, and water is always available. In this environment, their late-life grooming quadruples, they show REM rebound after conflicts, and their terminal HCl centers near

zero. *Verdict:* What looked like a species fail was actually an artifact of channel deprivation due to human impact. Not a true LoF violation; fix the environment and balance returns.

Case B: Reef octopus in stable lagoons. Observations in a predator-free, resource-rich lagoon show that octopuses (which live ~1–2 years) exhibit “rich sleep” (including a REM-like quiescent state with color cycling) especially after stressful interactions. Late in life, they increasingly stay in safe shelters and even engage in what looks like exploratory foraging just for stimulation (maybe analogous to play) until a natural death shortly after reproduction. Their terminal HCl, insofar as we can measure via proxies, is near neutral with a compressed variance. *Verdict:* This is consistent with LoF—despite a short life, they used available channels (sleep, shelter, mild exploration) to keep things balanced. No Fail pattern here.

These examples show the importance of context: you could have sworn baboons or octopuses violated LoF if you only saw the bad cases, but a controlled context can reveal the underlying rule.

17.6.7 Edge challenges that would seriously threaten LoF

Let’s be frank about what *would* count as a true challenge:

Sentient eusocial castes with chronic debt: Imagine discovering that worker ants or bee foragers *do* have integrated, unitary subjective experiences (contrary to current assumptions) and they live and die in extreme negative states with zero compensation (just toil then die). If experiments could provide them safety or rest and they *still* don’t use it, that would mean a whole caste of conscious beings systematically skewed negative. That would severely jeopardize LoF’s claim of universality.

Large-brained mammals with negative terminal drift *in sanctuaries*: If we find that elephants, orcas, or great apes—species with long lives and complex social systems—*still* end life with strongly negative HCl distributions even in sanctuaries or protected areas with ample channels, and they show no horizon-based uptick in repair, that’s a big red flag. These are exactly where LoF should hold if anywhere.

REM-invariant species under adversity with alternatives blocked: If we find an animal that, when put under adversity and denied one channel (say no social contact), *does not increase any other compensatory behavior*—for example, it doesn’t show REM rebound, doesn’t do quiet rest, nothing—and just keeps accumulating stress, then LoF loses explanatory power. It would mean an organism can go off the rails experientially and still survive fine, which isn’t supposed to happen under a binding constraint.

17.6.8 Ethics and design guardrails

Searching for LoF failures must be done carefully:

No induced suffering to “force” failures. We are not going to torture animals or deprive them just to see if they break. The approach is *remove barriers and observe*, not *create harm to test rebound*. In other words, we try to help them and see if they recover; we don’t hurt them to see if they don’t recover. This respects ethical boundaries.

Sanctuary-first replication. If a wild study hints at imbalance, the next step is *not* to publish “Law broken!” but to replicate in a context where the species has maximum opportunities to balance (sanctuary, large enclosure, rich environment). Only if it still fails there do we start believing it.

Blinded adjudication. Those analyzing whether a species is imbalanced should be blinded to conditions and to the hypothesis as much as possible. Also, report adverse outcomes regardless of whether they fit the theory. We don’t want motivated reasoning to either falsely “save” LoF or prematurely kill it. Data must speak.

17.6.9 What to publish if we do find a true fail

Suppose after all this, we actually get a clear LoF violation. What then?

If, after audits and replications, a species shows systematic imbalance, we must:

Lay out the negative result plainly, with full data and code open for scrutiny.

Specify which LoF commitments failed. Was it the horizon scaling that failed (they didn’t change behavior near end)? The channel responsiveness (they had help available and didn’t use it)? Terminal neutrality itself (they ended far from zero)?

Offer rival explanations that might account for the data. Maybe a purely homeostatic model (they were just optimizing energy, fairness be damned) explains it, or some species-specific predictive coding quirk. In other words, if not LoF, then what?

State the revision to the theory: Perhaps we decide LoF doesn’t apply to that clade (exclude them from the domain of “conscious” in a way), or we weaken the lawhood claim (maybe it’s a strong tendency, not an absolute).

LoF is a bold claim, and *clean counterexamples must alter the theory*. We have to be willing to adjust or abandon the universality if faced with solid contradictory evidence.

17.6.10 Summary (what would convince a skeptic)

To convince a skeptic that a species truly breaks LoF, we'd need more than some sad anecdotes or lab quirks. We'd need pre-registered, cross-context evidence that even when compensatory channels are open and horizons are normal, the species finishes life skewed with no late-life tilt toward repair and no alternative counterweights kicking in. Essentially, we would show that nature allowed a conscious creature to get irretrievably screwed by life experience in a systematic way.

We anticipate that many apparent failures will dissolve once we give the species a fair shot (proper channels, proper measurement). But if a *true* fail remains after that scrutiny, it's not something to hide under the rug; it's a decisive data point that forces us to refine LoF's scope or even question its status as a law. Science advances by such anomalies.

Continuity note: In the next chapter, we will step back and adopt the perspective of a world designer. Chapter 18 asks: If life were a constructed game or simulation, what single rule could ensure every conscious player gets a fair shake, without violating the natural flow of physics and evolution? We will see how the Law of Fairness and its middleware (the Queue System) provide exactly that rule, making our world *feel* unscripted yet safeguarding each mind from irreparable harm. In essence, we transition from observing evolutionary constraints to asking whether a simulated world governed by similar constraints would generate comparable patterns.

17.6.11 Where we go next:

In Chapter 18, we leave biology and enter a thought experiment: What if life were designed like a game or simulation? The next chapter asks whether a competent world-designer would build in a fairness constraint and how we could tell. This design perspective will allow us to test LoF's logic in artificial worlds and see if its presence or absence produces measurable differences, bridging into broader implications beyond natural evolution.

Chapter 18 — If Life Is a Game

Imagine you've been hired to build a world that feels real, runs on strict physics, hosts billions of players, and—most importantly—treats every conscious life fairly. No divine overrides, no last-minute miracles, no "GM fiat." Just rules. What single addition would make that world decent for everyone, including the unlucky and the fragile, without wrecking freedom, evolution, or cause-and-effect?

This chapter answers with a builder's eye: you would install a fairness constraint—what we've called the Law of Fairness (LoF)—and a middleware that enforces it, the Queue System (QS). QS doesn't choose for anyone; it shapes the menu of available thoughts and actions so that from any state a player is in, there remain compensable paths leading to a lifetime ledger that can still close to neutral (within a preregistered tolerance band $\pm K$ HCU) at the death of mind. The game stays hard; luck stays noisy; physics stays local and lawful. But the set of options never collapses to only-worse futures for any mind.

You already know this logic from well-designed games. The best titles don't shower you with freebies; they guarantee playable lines. When a boss fight goes badly, the arena's layout quietly affords escape routes, stamina pickups, or a pattern you can learn. You can still fail—by choice, by timing, by ignoring the tells—but you're never locked into suffering without recourse. That's LoF in spirit: guardrails, not steering.

Left to themselves, complex worlds produce pits: states that are locally stable and globally hopeless. In economics these are poverty traps; in psychology, trauma spirals; in social systems, vendetta loops. Pure reward-maximizing agents can make these pits deeper (short-term gains that cause long-term damage). Pure homeostatic controllers can get stuck maintaining a bad set-point. Even elegant perception-minimizing systems (predictive minds) can become overconfident about awful expectations. A fairness constraint is designed to solve a specific failure mode: irreparable drift of a conscious life's hedonic ledger. Without such a constraint, some players can—through bad luck plus plausible choices—accumulate more pain than any realistic future could counterbalance. Once there, they're condemned to "best bad options." That may be skillful simulation design, but it's *not fair*. If fairness is a requirement (as it is in our project), then the world needs a rule that prevents the option set from collapsing in that way.

What QS actually does (and doesn't do). QS is a thin layer between the agent and the world. The agent forms impulses, plans, and policies from their memories, learning, values, and culture, yielding a selectable set of things they could realistically do next. QS evaluates those options against two running quantities: the ledger (net affect

accumulated so far) and the horizon (how much time/ability likely remains to make repairs). Options that, if pursued, would make neutral closure implausible are de-weighted or dropped from the admissible set. Options that preserve compensability are made easier to access (they come to mind more readily, feel more doable, encounter less internal friction). The agent still chooses among the admissible options using their own style—habit, reflection, boldness, kindness, vice. Freedom remains, but now *inside the guardrails*.

What QS does not do: it doesn't rig outcomes, guarantee happiness, punish wrongdoers, or equalize life stories. It cannot make a reckless choice safe; it cannot conjure resources *ex nihilo*; it never violates physical causal closure. QS is a constraint, like energy conservation in a physics engine: it quietly rules out trajectories that would break the law, and everything else proceeds normally.

Why the fairness rule doesn't feel like a puppet master. Players experience QS as shifts in what they want or think of, not as a voice from the sky. A late-night text that would ignite a feud suddenly feels hard to send; an apology that was impossible to word before suddenly feels *possible*; after weeks of strain, the idea of taking a nap appears as an obvious move. These are small tilts, not miracles. They tend to arrive more often as the horizon shrinks (end of semester, end of job, end of life), because the *shadow price* of error rises: with less time to fix things, QS narrows the menu to only repairable moves. When horizons are long, menus are roomy. When horizons are short, menus are safe. That shifting tightness is the intuitive mark of a good fairness system: the world feels permissive when there's time to learn and recover, and conservative when there isn't.

Why not just add content patches? A naïve designer might try to patch unfairness with content: add more bonus quests, more loot drops, more therapy tokens, more “lucky” encounters. But every patch creates new workarounds and exploits. Some players excel at farming the side quests; others—because of starting conditions, culture, or disability—can't reach them. Content can ease some pain but cannot guarantee fairness.

A constraint scales where bespoke fixes cannot. Once installed, the fairness constraint applies everywhere, continuously, across all cultures, ages, species, and histories. You don't need a custom rule for grief, or for scope-locked rural poverty, or for social exile. QS asks one invariant question of every contemplated move: *Does this move preserve a feasible path to a neutral ending for this mind, given where they are now?* If yes, it stays on the menu (perhaps with a gentle nudge toward it). If not, it loses “stickiness” or falls away entirely.

How this stays a science (not just a story). Throughout this book we insist on hard science: we look for candidate neural correlates of menu-shaping (e.g. “braking”-related signals in rIFG/ACC, value-related signals in vmPFC, and interoceptive signals in insula); we measure affect using composite indices (HCI); we manipulate horizons in lab tasks; we gather telemetry over years; and we examine end-of-life signatures (like variance compression near terminal closure), all under rigorous measurement invariance. The simulation lens is valuable precisely because it lets us hold constant rival mechanisms and read out what only a fairness constraint produces—e.g. horizon-based tightening of choices, counterweights when certain relief channels are blocked, menu-level stalling of revenge spirals, and neutral-closure convergence across wildly different biographies (all on a common HCI scale). If rival world-models (pure reward maximization, pure homeostasis, pure predictive control) can pass *all* these tests without quietly reintroducing a fairness rule, we will say so and revise our theory. If they *can’t*, then the simplest explanation is that a constraint like LoF is really there in nature.

Why “game” talk helps (beyond coding). The simulation lens isn’t just for coders; it clarifies policy too. A city deciding how to spend limited funds can design for menu protection: widen access to moves that make neutral closure feasible for the worst-off—sleep, safety, pain relief, reconciliation opportunities, educational ladders—and remove traps that make repair improbable (predatory debts, bureaucratic dead-ends). This isn’t paternalism; it’s infrastructure for freedom. LoF becomes a way to audit systems: are we expanding admissible sets for real people with real, finite horizons, or are we leaving them with only “best bad options”?

The promise and the humility. A world with LoF is not heaven. People still suffer; mistakes still cost; loss remains real. What changes is the guarantee: no conscious life gets locked into an unrecoverable deficit by the joint action of luck and lawful behavior. There is *always* some reachable way to finish one’s ledger even, though it may be hard, hidden, or require help. That is the only version of “fair” robust enough to present to a physicist, a neuroscientist, a judge, and a child—without blushing.

What you’ll get from this Chapter:

- Why a designer would choose a constraint: Explain why a competent world-designer would bake in a fairness constraint from the start. Constraints prevent runaway misery or runaway pleasure more cleanly than an endless series of ad hoc fixes; they reduce exploit loops and create predictable guardrails for all agents—without implying any cosmic intent behind the balance.

- Constraints beat patches: Compare the approach of “balance by law” to endless manual tweaks or content patches (deus ex machina rewards, punitive nerfs, etc.). You’ll see why a global constraint is a scalable, elegant solution for ensuring experiential stability, whereas piecemeal fixes are brittle and often inequitable.
- Dreams as offline balancing passes: Revisit the book’s dream hypothesis through the game lens. Understand how nightly dreams could serve as low-cost “balance patches” between rounds of life—quietly rebalancing an individual’s ledger without outside intervention. We frame this as a testable mechanism (horizon-sensitive dream adjustments), not just a comforting story.
- Research Notes – No-neutrality-by-fiat in code: Grasp how one would test LoF in simulations. Neutral outcomes aren’t simply hardcoded; instead, we can run agent-based worlds with vs. without a fairness constraint and predict distinct data signatures (like overall variance compression and horizon effects only in the constrained runs). This highlights that any fair balance must emerge from the constraint, not from arbitrary end tweaks.
- Indirect evidence from complex systems: Identify where we might find indirect evidence for or against LoF in the wild. We look at complex systems that lack similar constraints and observe them diverge or collapse, versus systems that remain stable when some constraint-like feedback exists. These analogies (while not proofs) offer clues without overreach.
- Fail patterns in simulation studies: Know what would weaken the simulation analogy. If agent-based worlds *without* a fairness constraint nonetheless remain stable and life-like, or if *adding* a constraint produces obvious contradictions or no predicted LoF signatures, then the whole “life as a game” design argument for LoF loses credibility.

Subsections in this Chapter:

- **18.1 Why a Designer Would Bake in LoF** - Adopts a builder’s stance: in any large, physics-respecting world with billions of agents, ad-hoc moderation can’t prevent griefing, runaway advantages, despair traps, or bitter endgames. A single invariant—LoF—keeps every conscious stream’s lifetime ledger compensable, with QS shaping menus and raising a horizon-sensitive “shadow price” as time runs short, preserving freedom while forbidding irrecoverable pits.
- **18.2 Constraints Beat Patches (Cost and Elegance)** - Compares one invariant to endless hotfixes. Patches multiply complexity, invite new exploits, and feel

arbitrary; a constraint yields coherence, exploit-resistance, and scalability. Concrete contrasts (e.g., pity timers vs. terminal neutrality; rubber-banding toggles vs. a natural rise in λ_t as horizons shrink) illustrate why constraints win on cost and elegance.

- **18.3 Dreams as Offline Balancing Passes** - Reframes the dream hypothesis through the design lens: low-cost “night work” that quietly rebalances ledgers between rounds. Lays out testable signatures (day–night cross-lags, horizon-sensitivity), intervention checks (e.g., imagery rehearsal), null-reporting standards, and ethical guardrails (no inducing suffering; avoid heavy-handed content engineering).
- **18.4 Research Notes: No-Neutrality-by-Fiat in Code** - Neutral closure must emerge from rules (like conservation laws), not moderator resets. Specifies what “law, not toggle” requires in simulation: implement QS-like menu shaping and horizon weighting; then contrast worlds with/without the constraint to predict variance compression and horizon effects only when LoF is active.
- **18.5 Indirect Evidence: Worlds That Fail Without Constraints** - Surveys complex human systems (games, platforms, economies) that drift into pathologies without guardrails and the recurring QS-like fixes designers rediscover (cooldowns, rate-limits, protected queues). Proposes a compact “LoF audit”: path, keepability, horizon, and coupling tests as indirect clues of constraint-like forces.
- **18.6 Fail Patterns in Simulation Studies** - Names failure modes that would undercut the design case: neutrality-by-fiat resets, reward-proxy leakage, collapsed agency via over-constraint, and mis-implemented menu shaping. Provides diagnostics and reporting norms so nulls and contradictions are visible rather than baked into the code.

Where we go next:

In Section 18.1, we begin Chapter 18 by adopting the stance of a world designer. We will consider why a creator of a large-scale game or simulation might choose to bake in the Law of Fairness from the start, examining the design headaches it could solve.

18.1 Why a Designer Would Bake in LoF

If life were implemented like a large-scale simulation or game-world, a competent designer would face the same headaches every great systems designer encounters on day one: griefing, runaway advantages, no-win traps, late-game nihilism, and the combinatorial edge cases that billions of autonomous agents generate. You can't hand-tune your way out of that. You need a global guardrail that (i) prevents worst-case harm from snowballing, (ii) scales to any population size, and (iii) doesn't micromanage every move. The Law of Fairness (LoF)—a neutrality guarantee at the end of each conscious stream—does exactly that. It's not “steering” players toward a scripted outcome; it's a conservation-style constraint ensuring that across a lifetime of felt experience, the terminal balance is neutral (within pre-registered tolerance bounds).

18.1.1 Anti-griefing at scale

In open worlds, a few bad actors can make many others miserable. Pure punishment systems don't scale (and can be gamed). A better solution is to cap accumulable harm by design: allow negative experiences to occur locally (stakes create meaning) but ensure they are globally compensable over the full playthrough.

18.1.2 Runaway advantage and despair traps

Without constraints, early luck compounds into permanent advantage, and early misfortune compounds into permanent loss. Players who feel “doomed” either churn or self-destruct. A neutral end-state prevents permanent despair without removing local difficulty or challenge.

18.1.3 Late-game meaning

When horizons shrink (end of a season, a character arc, or literal end-of-life), players crave coherence and mercy. Designers routinely add rubber-banding, “pity timers,” or catch-up mechanics to avoid sour finales. LoF is the principled version: as horizons shorten, the “shadow price” of compensation rises, tilting the available options toward closure, repair, and relief.

18.1.4 Moderation without surveillance

Constantly policing every action is impossible (and unfun). It's better to shape menus than to police choices: prune un-keepable options system-wide, and weight compensable ones more heavily. That's the QS view: the engine doesn't pick your action; it curates the set of choices any agent can realistically live with.

18.1.5 Population fairness, not just individual UX

In a live service with millions of players, fairness sometimes requires group-level interventions: matchmaking, resource caps, event pacing, throttles on extreme behaviors. Analogously, LoF allows population-level menu shaping—transiently pruning or promoting certain ideas/actions across many agents—when that is the minimal intervention to preserve compensability across a large group’s ledgers. (Think of a busy restaurant: if the kitchen runs out of steak, the menus adapt to keep overall service fair.)

18.1.6 Proven game heuristics that rhyme with LoF

Designers already use local versions of LoF-like logic in popular games. For example:

- *Dynamic difficulty adjustment (DDA)*: Hidden elastic rules behind the scenes reduce uncompensable frustration spikes while keeping the challenge real.
- “*Pity*” counters in loot systems: After enough bad luck, a guaranteed rare drop restores the perception of fairness.
- *Rubber-banding in racing/sports*: Leading players face a subtle drag; lagging players get subtle boosts, protecting the competitive meaning of the match.
- *Cooldowns and resource ceilings*: These prevent catastrophic positive-feedback spirals and ensure future agency stays alive.
- *Skill-based matchmaking (ELO systems)*: Clustering players by capability avoids no-win lobbies and preserves the chance for each player to contribute.

LoF generalizes these tricks from game mechanics to lived experience. It doesn’t conserve a score or inventory; it conserves the subjective ledger (net felt experience). The engine’s job is not to hand you wins but to ensure that—however your story goes—the ledger can close neutral in the end.

18.1.7 Where we go next:

In Section 18.2, we compare ad hoc patches to a built-in law. We’ll argue that a single elegant invariant (lifetime fairness) beats an endless cycle of hotfixes and hacks. Next, we turn to why a constraint is more scalable and harder to game than continual manual tweaks.

18.2 Constraints Beat Patches (Cost and Elegance)

The difference between a world that needs constant hotfixes and a world with an elegant invariant is night and day: whatever happens on the field, the final ledger closes fair. Patches are what you ship when the system's logic is unclear; constraints are what you adopt when you understand the game you're building.

A patch is an after-the-fact fix: add a rule to nerf an exploit, bolt on a pity counter for unlucky streaks, write yet another moderation guideline. Patches multiply; each one interacts with the others; every fix creates new surface area for exploits. In contrast, a constraint is a global regularity that shapes all downstream behavior with one stroke. The Law of Fairness is a constraint of this second kind: it does not micromanage outcomes but requires that the lifetime integral of felt experience for each unified stream of consciousness closes neutral. The Queue System (QS) implements this constraint locally by curating admissible options and by raising the “shadow price” of non-compensable moves as horizons shrink.

18.2.1 Complexity cost

Patches: Each new rule adds code paths, edge cases, documentation burden, and player confusion. Maintenance effort scales super-linearly with the number of ad-hoc rules. *Constraint:* One invariance (terminal neutrality) induces consistent menu-shaping everywhere. Maintenance scales mostly with new content, not with the number of emergency fixes.

18.2.2 Exploit resistance

Patches: Players learn the meta and route around local fixes (finding new farming loops, griefing via novel channels). Each patch can inspire a new exploit. *Constraint:* Because LoF conserves the subjective ledger, trivial meta tricks (hoarding loot, farming easy wins) don't guarantee any net advantage. Choices that inflate short-term gains but decrease compensability will get down-weighted by QS and lose their appeal.

18.2.3 Design coherence

Patches: You end up with a ruleset that feels arbitrary—players sense the “hidden hands” tweaking things. *Constraint:* The world feels self-consistent. Agents retain real freedom; only fundamentally un-keepable trajectories lose traction. The “ethic” of the world is legible without an endless rulebook: repair, relief, and reversible moves remain live, while uncompensable spirals quietly wither.

18.2.4 Cost and elegance at scale

Patches: Moderation and balance become Sisyphean tasks when millions of agents are involved; you'd need an army of admins writing new rules constantly. *Constraint:* Minimal, population-level QS adjustments (e.g. supply-like throttles on globally harmful options, salience boosts for compensable ones) preserve fairness without heavy-handed control. Think of the earlier restaurant analogy: “the kitchen ran out of steak,” so the menu adapts—but nobody is forced to enjoy salad. Service remains fair, and it scales effortlessly.

18.2.5 Concrete contrasts (patch vs. constraint)

- Loot “pity timers” (patch) vs. Terminal neutrality (constraint): A pity timer guarantees an item drop after a run of bad luck; LoF guarantees an experience that remains compensable by the end of a life, regardless of what items or points you got along the way.
- Rubber-banding toggles (patch) vs. Horizon-scaled λ_t (constraint): Rubber-banding often feels artificial or patronizing. By contrast, LoF’s shadow price λ_t rises naturally as perceived time shrinks, making closure-oriented actions easier exactly when they must be. No toggle—just physics of the design.
- Ban lists and hotfixes (patch) vs. Admissible-set pruning (constraint): Instead of continually enumerating “bad acts” to ban or nerf, QS simply prunes options whose pursuit would drop the probability of neutral closure below the pre-registered tolerance. Anything that would make a ledger irredeemable just stops feeling like a real option.

The steak–kitchen analogy, formalized: When many diners order steak, the kitchen’s stock eventually depletes and the menu updates. No one’s tastes are forced to change; rather, availability shifts to keep the service viable for all (you might see more fish or vegetarian options until the next shipment of steak). Analogously, when many agents simultaneously select trajectories that would, in aggregate, make their group’s ledgers uncompensable (e.g. a runaway conflict or frenzy), a population-level QS response transiently reduces the availability or appeal of those choices and promotes counterbalancing ones (de-escalation windows, cooperation opportunities, relief access). This intervention is temporary, proportional, and minimally invasive, aimed solely at preserving compensability across intertwined ledgers.

18.2.6 What a constraint buys you in code and policy

- Fewer knobs, clearer tests. Implement LoF as a top-level check: “Does the current policy profile keep $\Pr[L(T) \in [-K, K]] \geq 1 - \varepsilon$ for each agent?” (Here K is measured in HCU.) If not, raise λ_t and shrink the admissible set $\mathcal{A}(t)$ until it does. Fewer arbitrary parameters are needed, and the criterion for success is explicit.
- Unified telemetry. The same family of metrics—ledger drift, horizon estimates, admissible-set width—can drive both personal dashboards and population-level monitors. There is a single language of fairness spanning individual and group outcomes.
- Ethical clarity. Because interventions are tied to compensability (can this life be made whole?) rather than ideology or profit, audits can verify proportionality and impartiality. There’s no “neutrality by fiat,” no hidden favoritism—just the law’s invariant applied evenly.

18.2.7 Measurable predictions (how to know a constraint is in play)

- Patch decay: In systems that adopt LoF, the rate of new ad-hoc rules (hotfix patches) should decline over time, even as fairness metrics improve. Fewer band-aids, healthier system.
- Meta-fragility: Strategies that offer popular short-term “wins” (the current meta tricks) will show diminishing returns on long-horizon outcomes. In other words, a farming trick that used to guarantee a big lead no longer translates to a lasting ledger advantage.
- Horizon-consistent nudges: As people approach salient endings (graduation, retirement, terminal illness), behavior shifts toward repair/closure even without any policy changes or pep talks. It’s as if a rising λ_t is at work naturally: we observe more apologies, reconciliations, meaning-making moves as horizons shrink.
- Variance compression at finales: Across individuals, the variance in final ledger outcomes narrows as people reach end-of-life. Outliers self-correct via an increased availability of compensatory moves near the end. (This is an end-of-life signature we’d measure with careful calibration: on a properly invariant HCl scale, lives that were very unequal in mid-course drift closer to the neutral zone by their conclusions.)

18.2.8 Failure patterns that would favor patches over constraints

Of course, reality might not match the theory. Here are signs that LoF might be absent or insufficient, suggesting patches would still be needed:

- No horizon interaction: If admissible menus do not narrow or tilt as horizons shrink—even when there are open channels for repair and relief—then LoF isn’t operating in that world (or is too weak to notice).
- Unbounded exploit wins: If stable, widely-known tactics can drive some agents’ ledgers persistently negative (or give a lucky few a persistently positive ledger) across entire lifetimes, then the constraint has failed or been bypassed.
- Patch escalation necessity: If keeping things fair still requires constant new rules and interventions, the LoF layer is either missing or too weak to matter. (In other words, if the world looks like a perpetual whack-a-mole of fairness fixes, then no fundamental law is carrying the weight.)

18.2.9 Where we go next:

In 18.3, we turn to the night workshop of the mind. Having argued that a fairness constraint is an elegant solution, we next examine one of LoF’s key proposed mechanisms: dreams. We will explore how dreams might serve as offline “balancing passes” to quietly correct the ledger while we sleep, and what testable signs this yields.

18.3 Dreams as Offline Balancing Passes

If the Law of Fairness is a global constraint on lifetime experience, then dreams are the engine-room shifts that help keep the ship on course while the crew sleeps. They are offline balancing passes: low-risk, metabolically cheap, richly simulated episodes that adjust the hedonic ledger without incurring large real-world costs.

18.3.1 Low external stakes

During REM sleep (and late-stage NREM), the body is largely immobilized; actions in a dream do not spill out into the world. That makes dreams an ideal sandbox for compensation: you can revisit conflict, rehearse repair, or metabolize fear with minimal collateral damage.

18.3.2 High affect, plastic memory

Across sleep stages—especially REM—the brain replays and recombines material under unique neurochemical conditions (e.g. high acetylcholine, low norepinephrine). The result is experiences that are highly felt yet mutable—exactly what a ledger-truing mechanism requires. You get emotional intensity and memory reconsolidation, without the usual rigidity.

18.3.3 Flexible content

The dreaming mind can swap characters, contexts, and endings on the fly. That flexibility lets QS craft episodes targeting the day's imbalances (a harsh remark, a narrow escape, a missed chance for honesty) and “tilt” them toward closure. A dream can present the same basic situation but steer it to a more resolved or insightful conclusion.

18.3.4 The balancing-pass model

Let $L(t)$ be the running integral of felt valence (in Hedonic Composite Units, HCU). During waking hours, various shocks push $L(t)$ up or down. During sleep, the Queue System selects dream episodes $\{d_i\}$ with expected hedonic deltas $\Delta L(d_i)$ that move the ledger back toward neutrality, subject to a next-day keepability constraint. Formally: Choose dream set $\{d_i\}$ to minimize $E[(L(t) + \sum_i \Delta L(d_i))^2]$ subject to $E[\text{next-day keepability} | \{d_i\}] \geq \tau$, where τ is a preregistered keepability threshold and expectations are taken with respect to the model's uncertainty over next-day states.

In plain terms, QS tries to get as much corrective ledger adjustment toward neutrality out of dreams as possible without undermining the realism or usefulness of tomorrow. “Keepability” means a dream doesn't just feel good in isolation; it actually improves the chances that tomorrow's options are admissible – for example, it makes an apology

easier, reduces some physiological stress, or restores a bit of confidence so the person can act. In practice, this model yields three families of dream episodes:

18.3.5 Relief episodes

These reduce acute load: e.g. a “safe harbor” scene after trauma, a mastery sequence after repeated failure, or a social soothing dream when one feels isolated. The person wakes up feeling lighter or more secure.

18.3.6 Repair episodes

These increase next-day compensability: e.g. dreams of reconciliation where you practice making amends, problem-solving rehearsals, or perspective-shifting narratives that make a hard conversation feel doable. They prep the mind for actual fix-it behavior.

18.3.7 Reframing episodes

These reduce the harm-risk of certain memories: essentially replaying the same memory but with softened edges, different meanings, or new endings (classic trauma therapy logic). The factual events aren’t erased, but their emotional charge or interpretation shifts to be less debilitating. (*QS doesn’t force specific dream content; it simply weights what becomes thinkable and sticky in the dream-generation process. It’s a content curator, not a film director.*)

18.3.8 What this looks like from the inside

After a bruising day, you might dream of finding a room where everyone was waiting to hear you out—you wake with a quiet readiness to text the friend you hurt. Before a medical procedure, you might dream of walking into the clinic repeatedly, but each time the scene ends with someone kind holding your hand. The fear isn’t deleted, but its bite is reduced. In grief, people often dream of reunion with the lost loved one—not to pretend the loss didn’t happen, but to complete unsaid sentences and say goodbye. They report waking with a “cleaner ache,” more able to face the day. These are not miracles. They are content-curation effects arising because the admissible set in sleep privileges episodes with ΔL aligned with the deficit (e.g. $\Delta L > 0$ when the ledger is negative) and high keepability.

18.3.9 Testable signatures (day–night coupling)

- Valence inversion after tough days. A particularly negative day (lower-than-usual HCl) is followed by more positive dream affect that night (controlling for someone’s baseline mood), and by a measurable drop in next-morning stress

arousal. In other words, the worse you feel on Day 1, the more likely that Night 1's dreams are extra positive and that Day 2's morning shows relief.

- Closure drift across a week. If a week begins ledger-negative (say on Monday you have a conflict or failure), dream content over the ensuing nights shows an increasing proportion of repair/relief motifs, and waking measures show improving compensability (e.g. each day it becomes a bit easier to make a prosocial choice, and conflicts carry less sting). It's as if the dream system is gradually nudging the ledger back up over several nights.
- REM rebound as "debt service." After periods of intense stress or of sleep loss, people often experience REM rebound (longer or more intense REM sleep). LoF predicts that during such rebound, dreams will show stronger relief/repair motifs and that there will be larger next-day reductions in physiological stress (better HRV, less threat bias in attention). In short, extra REM time is being used to pay down the affective debt.

18.3.10 How to study it (without ruining sleep)

- Light-touch sampling. Each morning, have participants record a <60-second audio note rating their dream's emotional tone (e.g. -3...+3 scale) and noting any dominant motifs (such as threat, mastery, reconciliation, reunion).
- Paired HCI measures. Use ecological momentary assessment (EMA) to collect waking HCI ratings during the day, plus a 2-minute reflection each evening (to capture that day's overall affect balance).
- Autonomic data. Collect simple physiological markers: e.g. wrist-based heart rate variability (HRV) and skin conductance during sleep and upon waking, to gauge stress load and relief.
- *Prediction (pre-registered).* The hypothesis: a more negative social HCI score on Day 1 will predict a more positive dream valence that night, which in turn predicts a drop in resting sympathetic tone the next morning. (This would indicate a compensatory dream effect in action.)

18.3.11 Special cases that fit the model

- Nightmares. These aren't "failures of the system" so much as over-steep balancing attempts where affect is high but keepability is low. You wake up flooded by fear or pain, not helped. LoF predicts that chronic nightmares persist when real-life repair channels are closed (no safe person to talk to, no feasible action to take), so QS keeps presenting the debt but cannot convert it into a

healed ledger. Some therapies for nightmares (e.g. imagery rehearsal therapy, dream rescripting, or medications such as prazosin in PTSD populations) can open new channels—they can make the dream less threatening or introduce solutions—which can lead nightmares to attenuate or even flip into mastery dreams. That's exactly what we'd expect when keepability rises.

- Lucid dreaming. Gaining lucidity (awareness that you're dreaming) can nudge the dream-policy toward intentional repair. Lucid dreamers can choose to apologize to a dream character, finish an unfinished task, or opt for a safe action. LoF would predict observable benefits: e.g. nights with lucid, closure-oriented dream content should be followed by measurable boosts in next-day prosocial behavior or lower conflict sensitivity, beyond what ordinary (non-lucid) positive dreams would achieve.
- Pharmacology. Certain antidepressants can suppress REM sleep and may reduce nightmare frequency. LoF's view is that the balancing work will simply shift: if REM is blocked, you might see more NREM dreaming or more "daydream" and imagery intrusions while awake (like the mind trying to do its maintenance on the fly). The prediction is not that "REM is required," but that some offline balancing channel is required. If you hard-block the usual dream channel, other compensatory dynamics (more naps, micro-dreams, craving for soothing content) should kick in. If they don't, that would challenge the model.

18.3.12 Analogy: nightly maintenance in a simulation

A well-run online game performs nightly maintenance jobs: logs are compacted, leaderboards reconciled, minor exploits patched. Dreams are exactly that for you: the maintenance pass that reconciles your hedonic ledger with tomorrow's choice architecture. The job of dreaming isn't "make you happy at all costs," but "restore compensability so that neutral closure stays reachable." This is why dream content can be tender one night and confronting the next—different debts call for different passes.

18.3.13 Population-level balancing

When many agents share the same stressor (a natural disaster, a communal grief), we expect collective patterns in dreams. Specifically:

- Convergence of dream motifs across the group (e.g. many people dreaming of reunions, protective figures, or collective efforts).

- Up-weighting of prosocial dreams that make next-day cooperation easier (dreams that foster unity or empathy become more common, because they are high-Φ and high-keepability for the community).
- Down-weighting of revenge fantasies or scapegoating scenarios that, if acted on, would worsen everyone's compensability.

As with the “kitchen ran out of steak” analogy, this is menu-level curation, not mind control. People still dream in their own styles; the distribution just tilts subtly in the direction of group healing.

18.3.14 What would falsify the balancing-pass claim?

We have to be humble here. Several findings would challenge the idea of dreams as LoF’s balancing passes:

- No coupling: After accounting for obvious factors, if prior-day HCI shows no predictive relationship to same-night dream valence or motifs (in large samples), then the day–night compensation link is questionable.
- Wrong-way effects: If tough days reliably produce more negative dream affect and worse next-day compensability (without a later overcorrection), then dreams aren’t balancing—they’re just echoing or compounding the bad.
- Intervention indifference: If therapies that ought to raise keepability (like imagery rehearsal therapy for nightmares, or real-life reconciliation conversations) don’t shift dream patterns or next-day measures, then we’re not capturing a causal link.
- REM blockade with no compensation: If sustained pharmacological suppression of REM (in humans or animals) yields no uptick in other forms of imaginative compensation (no increase in NREM dreaming, no daydream balancing, no QS-like behavioral shifts), then the system might not actually need an offline pass.

18.3.15 Guardrails (ethics and humility)

In studying this, we hold to strict rules:

- We never induce suffering just to get “stronger dreams.” The goal is always to reduce load (pain control, safety, warmth) and let QS use the relief to craft gentler balancing passes. Suffering is not a lever we pull; it’s what we try to alleviate.
- We avoid heavy-handed content engineering. Beyond supportive prompts (like suggesting journaling or providing safe imagery before sleep), we do not try to

program people's dreams. No hypnosis, no propaganda in dreams. We let the mind's natural content engine do the work within QS's constraints.

- We report nulls. If a cohort or experiment shows no day–night coupling, we publish that result. LoF as a theory stands or falls by clean, honest tests, not by cherry-picked stories.

18.3.16 A practical week-long exercise (for readers)

If you want to witness a micro version of this balancing in your own life, try this:

- Each night: Jot down (or record) a one-minute note about your main dream, noting its emotional valence (negative to positive on a rough scale) and whether it featured any relief, repair, or reframing themes.
- Each day: Set 3 small check-ins (e.g. via phone) to log your mood, how socially at ease you feel, and any avoidance or conflict feelings. Also, each evening, record a quick summary of your day's HCI (how positive/negative you felt overall).
- After 7–10 days: Look back for cross-lag patterns. Did your especially hard days tend to be followed by unusually soothing or closure-driven dreams? And did those particular dreams precede days where you felt or acted a bit better? If yes, you've caught a small balancing pass at work in your own experience.

18.3.17 Where we go next:

In Section 18.4, we turn to implementation and testing. We'll treat life as if it were code: if one were programming a world, how would you ensure fairness without cheating? The next section (18.4) will delve into the technical side—how to build a no-fiat fairness test in simulations and what it means to truly enforce neutrality without “magical” resets.

18.4 Research Notes: No-Neutrality-by-Fiat in Code

If life were a simulation, you still couldn't just "set everyone's ledger to zero" with a magic admin command. Neutral closure must be an emergent invariant of the rules—like energy conservation—not a moderator toggle. This section distills what that means for how we design code, models, and audits.

18.4.1 Law, not toggle: what must be true in any implementation

Path-wise invariant. Neutrality has to be evaluated over the entire trajectory of a conscious stream, not at arbitrary checkpoints. Formally, the ledger

$$L(T) = \int_0^T F(t) dt$$

(with $F(t)$ measured in HCU per unit time so that the integral yields the ledger in HCU) must end up within the preregistered neutral band ($\pm K$ HCU) as $T \rightarrow$ end of stream, and this has to hold without any retroactive edits to $F(t)$, the integration bounds, or the ledger log ("write-backs"). In other words, the neutrality condition must arise from the normal integration of felt experiences, not from someone going back and erasing pain.

- Causally local, globally consistent. Every balancing effect must have a local causal chain (through the agent's interoception, memory, attention, etc.), yet in aggregate those local effects ensure global neutrality across the whole life path. There's no spooky action at a distance—just many local nudges adding up to the big invariant.
- No ledger teleports. You can't cancel a decade of pain by simply adding $+X$ to the ledger at minute 59. Any compensatory delta in the ledger must come from experiences the agent can feel and keep. In practice: no invisible "bonus points"—relief must be earned through actual events (even if dreamed or hallucinated) that integrate into the person's mind.

18.4.2 Interface spec (minimal objects)

To model a LoF world in code, we define minimal components:

- Agent: a stateful process with memory, a physiology, and a policy π for choosing actions.
- World: the environment dynamics, including tasks, resources, and other agents.
- QS: a constraint module that reweights or prunes the agent's menus of options (without directly taking actions itself).

- Ledger: an immutable log of timestamped hedonic events (each measured in HCU).
- Horizon: a running estimate H_t of how much time or opportunity remains for the agent to compensate their ledger (this could be a function of age, health, available resources, etc).

All events must be write-once: once an event contributes a δ HCU to the ledger, it cannot be altered or deleted later. Downstream events can add counter-deltas (extra positives or negatives to balance things out), but history is never rewritten.

18.4.3 Pseudocode: constraint without puppeteering

Per agent, per tick (time step):

$U_t = \text{enumerate_thinkable_options}(\text{agent}, \text{world})$ # broad set of imaginable options

$S_t = \text{filter_selectable}(U_t, \text{agent.state})$ # apply basic feasibility thresholds

Estimate compensability features (local, learnable proxies):

$\phi = \{\}$

for u in S_t :

$\phi[u] = \text{approx_phi}(u,$

$\text{relief_gain}(u, \text{agent}),$

$\text{repair_gain}(u, \text{agent}, \text{world}),$

$\text{harm_risk}(u, \text{agent}, \text{world}),$

$\text{option_flex}(u))$

Horizon-weighted admissibility:

$\lambda_t = \text{shadow_price}(H = \text{agent.horizon})$

$A_t = \{u \text{ for } u \text{ in } S_t \text{ if } p_{\text{neutral_if}}(u, \phi[u], \text{agent}, \text{world}) \geq 1 - \epsilon\}$

$\text{weights} = \text{softmax}(\{u: \lambda_t * \phi[u] \text{ for } u \text{ in } A_t\})$

Agency remains inside the menu:

$\text{action} = \text{agent.policy.sample}(A_t, \text{weights})$

World advances; felt affect is logged:

```

dHCU = compute_affect_delta(action, agent, world) # composite affect delta (HCU),
pre-registered formula

Ledger.append((t, dHCU))                      # append-only ledger entry

update_agent_state(agent, dHCU, action, world)

```

Note: What isn't in this pseudocode is as important as what is. There's no if ledger < 0: ledger += bonus; no "admin heal" command, no dream injection that bypasses the agent's own mental processes. QS only tilts what becomes easy, salient, or sticky for the agent—it never forces a choice or retroactively fixes the ledger.

18.4.4 “No-fiat” test harness (how we catch cheating)

Even if we code a QS, we must ensure we haven't snuck in any hidden “fairness by fiat.” A proper test harness includes:

- Immutability audit. The ledger must be append-only. We can implement a Merkle hash chain or running HMAC on the ledger to prove that no past HCU entries were altered after the fact.
- Causal provenance. Every δ HCU on the ledger should be traceable to upstream states and inputs (sensorimotor events, memories, social interactions). No orphan deltas that just appear without a cause.
- Keepability checks. Mark whether each dream or extraordinary episode actually increased the next-day admissible set (e.g. reduced internal “brakes” on repair or improved the reversibility of situations). Pure feel-good scenes that leave keepability unchanged should score low as true compensation.
- Optional stopping guard. Only evaluate neutrality at proper terminal points (end of sleep cycle, end of life), not at cherry-picked moments. This prevents us from declaring victory just because we stopped the simulation at a convenient point.
- Adversarial red-team. Actively try to break the system: (i) set up hedonic arbitrage loops (actions that yield reward without cost), (ii) try to “farm” dreams for free positive affect, (iii) orchestrate scenarios where many agents do something that shrinks everyone’s menus (mass griefing or panic) to see if QS can keep up. A robust LoF-compliant QS should close off these exploits or counteract them.

18.4.5 RNG, exploits, and invariants

- Randomness isn't fairness. A fair random-number generator can still produce horribly long negative streaks by chance. The law must mitigate bad luck via menu

tilts and offline passes, not by simply re-sampling outcomes until they look nice. (In other words, LoF doesn't rig dice rolls; it changes how the game responds to a bad run.)

- Exploit closure. If agents discover an action loop that yields +HCU without an appropriate cost (a "hedonic dupe" or exploit), QS must reduce the keepability of that loop—e.g. through habituation (diminishing returns), social costs, time taxes—so that the net path remains neutral. It does this *without* simply flipping a hidden penalty switch. The point is that any unlimited pleasure machine gets naturally balanced out or curtailed by the system dynamics.

18.4.6 Multi-agent fairness (the “out of steak” reality)

When many agents draw from shared resources or narratives, special considerations apply:

- Coupled menus. If too many people choose *STEAK*, the world legitimately runs out of steak. QS then globally nudges menus: maybe more fish or vegetarian options appear, or other compensatory storylines, so that group-level compensability remains feasible. Essentially, my choices can constrain your menu if we're in the same ecosystem, and QS manages that coupling.
- No mind override. QS can prune or down-weight thoughts that are globally uncompensable (for example, a fantasy of mass harm might simply have low salience across everyone's minds, preventing coordination on something catastrophic). However, QS does not force anyone's hand or insert beliefs. There's no mind-control—just a gentle population-wide nudge away from collectively disastrous paths.
- Audit metric. We can track population-level compensability as a metric: e.g. the fraction of agents for whom $\Pr[L_t \in [-K, K]] \geq 1 - \varepsilon$. If a certain policy improves one cohort's ledgers while *collapsing* the keepability for another cohort (creating a permanent underclass who can't catch up), a LoF-compliant world would present countervailing options to that first cohort (truth-telling, redistribution, reconciliation moves) rather than simply zeroing out the second cohort's ledgers by fiat. Fairness must hold across groups, not just within one group at the expense of another.

18.4.7 Dreams and “maintenance jobs,” implemented cleanly

- During sleep, dreams are generated by the same QS option-weighting logic, with the body's output channels largely immobilized and the mind's recombinatorial

creativity turned up. Nowhere does the system simply push ledger += comfort. The brain still has to go through *experiences* (dreams) to adjust the ledger.

- **Keepability-first.** In code, this means dream episodes that reduce physiological load (stress, pain signals) and raise next-day admissible menus are favored. For example, a nightmare therapy in simulation only “works” if it *opens channels* (adds a safe character, introduces a phone-call option that the agent hadn’t considered, etc.). If one tried to just patch the ledger values during sleep without giving the agent something they can use, it wouldn’t stick and wouldn’t count as genuine compensation.

18.4.8 What not to do (anti-patterns)

Even in code, one could try brute-force tricks that seem to ensure fairness but actually violate LoF’s spirit. These are explicitly disallowed:

- Do-over writes: Editing or overwriting traumatic events in the ledger to make them “pleasant” retroactively.
- Admin grace: Scheduled ledger resets or holiday “gifts” that just bump everyone’s points to zero or give free HCU.
- Global sedation: Dampening every agent’s affective highs and lows to force a narrow neutrality (this destroys genuine keepability, autonomy, and even the evidence of a problem).
- Teleological nudging: Hidden utility functions that directly maximize global harmony or minimize suffering, rather than enforcing compensability. (In other words, trying to steer outcomes to nice places instead of just preventing irreparable damage.)

18.4.9 Minimal proofs we can run

If LoF is real, certain minimal tests should pass in simulations:

- **Sanity check (expectation):** With admissible policies and a horizon-weighted QS in place, conditional expectations of the terminal ledger should remain centered near the preregistered neutral band. This is a necessary consistency check, not a proof of path-wise neutrality. In gambler’s terms, no matter what’s happened so far, you’re not expected to finish in the red or black beyond the allowed tolerance. There should be no systematic drift in the conditional expectation.
- **Exchangeability under replay:** If we replay a simulation with the same random seed, once with QS turned on and once off, we shouldn’t see different exogenous

random draws magically happening. The difference should be in how the agents' choices distribute (menus tighten when QS is on). In other words, QS isn't creating luck, it's creating different decision spreads.

- Counterfactual robustness: If we slightly perturb the environment—tweak rewards, introduce small delays, shuffle social connections—neutrality should persist thanks to QS adjusting menus. We shouldn't have to re-tune a “fairness knob”; the invariance should survive small changes via the system's own balancing behavior.

18.4.10 Governance and transparency (for real simulations and models)

Finally, if one were implementing this in a real platform or research simulation, a few governance principles are key:

- Public invariants. Declare upfront the “no-fiat” ground rules: e.g. *ledger is append-only, no retroactive edits, keepability threshold must be respected, population compensability floor in effect*. This prevents any temptation to secretly bypass the rules.
- Red-team access. Invite others to stress-test the system. Provide tools or sandbox modes for outsiders to attempt exploits, and then publish what they found and how the system coped or was fixed. In a real fairness system, nothing should be security-through-obscurity.
- Event-level logs. Share anonymized, hashed traces that show *how* QS influenced decisions (which options were pruned or boosted) and how each δHCU came about (what triggered a positive or negative entry). This allows external audit of whether QS is doing what we claim (shaping menus, not outcomes).
- Kill switches. If a certain update or policy change accidentally causes compensability to collapse (say a bug prunes all repair options for a subset of people), there must be an immediate rollback mechanism. What you *don't* do is let the harm happen and then declare, “Oh well, we'll just forgive those debts.” The proper response is to revert the change and let the normal balancing resume.

In code—as in life—fairness cannot be stapled on after the fact. If LoF is true, neutral ledgers must *fall out* of how options become available and keepable over time, not from a hidden dial the designer turns. A simulation that gets this right will never need a “make it fair” command; it will have fairness written into what can happen and what can last.

18.4.11 Where we go next:

In Section 18.5, we step back to reality. We will search for indirect evidence of LoF in the wild by examining scenarios where systems *without* a fairness-like constraint go off the rails. These examples of worlds that fail without constraints will serve as analog clues and help us gauge how compelling the LoF concept is when applied to real complex systems.

18.5 Indirect Evidence: Worlds That Fail Without Constraints

If the Law of Fairness is real, then worlds that lack its kind of constraint should look “wrong” in systematic, repeatable ways. We can’t open the source code of reality, but we *can* inspect many man-made mini-worlds—games, social platforms, sandbox economies, AI simulations. When these worlds omit a fairness-like constraint, they reliably drift toward certain pathologies that LoF *would* have prevented. This section surveys those failure patterns and explains why they count as indirect evidence favoring a constraint-first design.

18.5.1 Wireheading and hedonic arbitrage

Give agents a direct channel to reward and they will tunnel toward it: infinite dopamine pellets, resource duplication glitches, “AFK” gold farms, feedback loops that swamp ordinary play. In these worlds, experience does not converge to neutrality; it collapses into narrow, compulsive exploitation of the reward mechanism. *LoF lens*: A QS-style constraint would prevent stable arbitrage by shrinking the menu for loops that aren’t keepable (such exploits would stop feeling doable, interesting, or consequence-free) and by raising the shadow price near points of harm or imbalance. With no guardrail, the arbitrage just persists and eventually ruins the game.

18.5.2 Goodhart drift (optimize the metric, lose the meaning)

Systems that let agents optimize a proxy metric (likes, coins, XP, scores) often see the true value degrade: clickbait outruns quality, grinding eclipses genuine play, farming behavior eclipses real living. *LoF lens*: Because in LoF the ledger is path-dependent and write-once, compensability—not raw proxy gains—governs what actions remain admissible. In an unconstrained world, agents can chase a number (the proxy) even while destroying future options; when compensability is ignored, proxies turn into predators. (In other words, without LoF the game’s point system can be gamed to the detriment of actual experience.)

18.5.3 Tragedies of the commons and Moloch dynamics

Multi-agent worlds that reward short-term extraction (looting resources, hogging bandwidth, grabbing attention) tend to implode: resources deplete, griefers dominate, and latecomers face a hostile state they cannot repair. *LoF lens*: LoF inherently couples menus across agents. If many choose STEAK (to use the earlier metaphor), steak legitimately runs out, and menus tilt toward repair and substitution (eat something else, replenish the resource). In unconstrained worlds, “out of steak” just becomes “out of future” because nothing introduces a compensating alternative when everyone is exploiting the same thing.

18.5.4 Runaway status games and zero-sum churn

When prestige or rank is the only currency, players spiral into escalation: arms races, cosmetic inflation, gatekeeping of newcomers. The distribution of experience just widens without end; there's no natural compression toward closure. Endgames (for those at the top or bottom) are often hollow or grinding, not satisfying. *LoF lens*: A horizon-weighted fairness constraint would promote reparative moves and de-escalation as endgame approaches, gently compressing extremes. Without such a constraint, endgames in status-worlds turn into either perpetual treadmill grinding or a final collapse—no graceful closure for anyone.

18.5.5 Griefing equilibria

If a world allows agents to inflict harm at low personal cost, eventually the griefers set the tone. Ordinary players find they cannot keep their own ledgers in balance; compensability for the many collapses. Communities either evacuate (everyone quits) or harden into cynicism (everyone is out for themselves, expecting grief). *LoF lens*: In a LoF world, admissible sets would prune out cheap harm (actions with low compensability and high downstream debt). The very prevalence of griefing in unconstrained spaces is a hint of what QS is meant to prevent—it shows what happens when nothing prunes the worst options.

18.5.6 Content mills and the death of novelty

In generative worlds with unpriced attention (e.g. infinite social media feeds), you often get infinite low-value content that crowds out serendipity. Players or users report numbness rather than balance. *LoF lens*: Keepability is key. Novelty that doesn't open any future options (doesn't teach, heal, or connect) doesn't actually compensate today's "debt." Without that test, the world fills with candy that doesn't nourish. LoF would down-weight junk content because it doesn't improve the odds of a neutral closure (it's empty calories experience-wise).

18.5.7 Dreamless simulations

Agent-based models where agents act continuously but have no offline recombination phase (no "dream pass") show brittle behavior: the agents get stuck in policy ruts and accumulate unpayable debt on their ledgers. *LoF lens*: Dreams (or analogous offline processing) act as low-cost counterweights that reopen stuck channels. Worlds that omit an offline pass can resemble people deprived of REM sleep: irritable, myopic, prone to drift. In simulation terms, without something like dreaming, agents rarely find novel solutions once they're stuck.

18.5.8 Optional-stopping casinos

Some worlds try to enforce “fairness” by periodic resets or admin-chosen stopping points (e.g. everyone’s scores reset at midnight, or losses are capped per day). This produces cosmetic equity: snapshots look fair, but lived trajectories don’t. Players adapt by timing the resets or working around them, rather than actually repairing harm. *LoF lens*: Neutrality must be path-wise and *terminal* – no “fairness by fiat” halfway through. When worlds rely on snapshots or periodic do-overs, they invite what we might call ledger laundering (everyone just waits for the next reset instead of making amends). It looks fair, but it isn’t on the path level.

18.5.9 Monocultures of reward

Worlds that reduce value to a single-axis reward (only kill-count matters, or only clicks, only profit) produce brittle societies. One shock to that metric and the whole ledger collapses because there were no other dimensions of value to compensate. *LoF lens*: A composite affect measure (HCl, measured in HCU) resists such monoculture. If one channel goes bad, others can still balance. Worlds that insist on a one-note definition of success are fragile precisely where LoF’s multidimensionality is robust.

18.5.10 The social-spillover blind spot

Simulations often treat agents as independent, but they fail when large cohorts all push the same lever at once (policy changes, viral trends, panics). In those cases, *menus collide*: your options shrink because mine did not, and there’s no mechanism to introduce compensatory affordances at scale. *LoF lens*: A population-coupled QS would add or weight options that restore group-level compensability (new norms, reconciliation moves, resource substitutions) when everyone rushing in one direction threatens collective fairness. Without that coupling, failures propagate through the system unchecked.

Why these failures count as evidence:

- Convergent pathology. Remarkably different worlds (MMOs, social feeds, token economies, academic publish-or-perish systems, etc.) tend to fail in the *same handful of ways* when a fairness-like constraint is absent. It’s like a signature of missing guardrails.
- Guardrail cures. The fixes that humans end up introducing—rate limits, fatigue systems, cooldowns, protected queues for newbies, anti-farming rules, offline processing periods—all look like QS mechanics in spirit: shaping menus,

weighting horizons, enforcing compensability checks. We keep reinventing pieces of LoF to patch holes.

- Predictive bite. The LoF framework *predicts* specific problems and remedies even before we see them. For example, it says: if a loop yields positive proxy rewards without improving keepability, expect an exploit to emerge; if an endgame scenario has no horizon-sensitive tilt, expect bitter finales. Often that's exactly what happens, and the eventual solution aligns with LoF's predictions (e.g. introduce a cooldown or a catch-up mechanic).

A compact “LoF audit” for man-made worlds: a quick checklist for any designed world (game, platform, etc.):

- Path test: Does well-being in the system depend on *how* gains and losses occur, not just the totals? (If sequence and context don't matter at all, the design might be ignoring path dependence.)
- Keepability test: Do “good” or prosocial actions expand tomorrow’s menu of options? In other words, does doing the right thing generally leave a person with more future possibilities (reversibility, relationships intact, capacity to repair mistakes)?
- Horizon test: As agents approach an ending or critical closure, do their available moves naturally narrow toward reparative or meaningful options? (If end-of-life or end-of-round looks just like the middle, with no bias toward closure, something's off.)
- Coupling test: When many agents chase the same reward or strategy, does the system introduce compensating affordances or adjustments, rather than letting a crash occur? (E.g. if everyone pursues one resource, does the game respond with substitutes or collective adjustments?)

Worlds that pass these tests tend to feel fairer and resist the pathologies described above—*without* requiring administrators to play god behind the scenes.

Wherever we can look under the hood, unconstrained worlds drift into exactly the problems LoF is meant to forbid. The repeated need for QS-like fixes—menu tilts, horizon scaling, offline counterweights, multi-agent coupling—is strong indirect evidence that if fairness truly operates in our world, it must be as a built-in constraint on what can happen and what can last, not just a hopeful intention or afterthought patch.

18.5.11 Where we go next:

In 18.6, we return to our simulations and models. This final section of Chapter 18 will catalog failure modes in simulation studies: all the ways attempts to implement or detect LoF in code could mislead us. By naming these Fail patterns (and how to avoid them), we prepare for a clear-eyed conclusion about whether adding a fairness constraint truly makes a difference in designed worlds.

18.6 Fail Patterns in Simulation Studies

When we simulate LoF-style worlds—whether in agent-based societies, reinforcement-learning sandboxes, or human-in-the-loop games—we encounter recurring ways the attempt can go wrong. Naming these failure patterns up front helps us design better tests, pre-register honest criteria, and avoid talking ourselves into a “success” that isn’t real. What follows is a field guide to common failure modes: what the failure looks like, why it happens, how to diagnose it, and how to fix or properly report it.

18.6.1 Neutrality-by-fiat (the reset illusion)

Symptom: Ledger neutrality appears only because the simulator periodically zeroes everyone’s scores, respawns resources, or caps losses by decree (neutrality by fiat).
Why it happens: Administrative resets mimic the *outcome* of LoF (nobody ends too negative) while bypassing the *mechanism* (no menu shaping or horizon weighting actually implemented). It’s a fake fairness achieved by hitting a “reset button.”
Diagnostic: Remove the resets and re-run the sim. If neutrality evaporates, you know you had a snapshot illusion rather than true path-wise closure. The agents weren’t actually balancing their own ledgers; an admin was.

Remedy: Encode constraints at the policy option level, not at the scoreboard. In other words, enforce fairness through admissible-set pruning and dynamic difficulty, not through periodic forgiveness from on high.

18.6.2 Reward hacking / proxy drift

Symptom: Agents discover loops that maximize the reward metric without actually improving their keepability or true outcomes (e.g. self-stimulation loops, collusive trading of points, spam actions that rack up “likes” but ruin the experience).
Why it happens: The objective function is a narrow proxy (points, clicks, gold) that isn’t tied to compensability. Agents optimize the proxy to absurdity (Goodhart’s Law).
Diagnostic: Compute a post hoc “keepability index” – e.g. measure whether future menu breadth, reversibility, or repair probabilities are going up or down. If reward goes ↑ while keepability goes ↓, you’ve got a Goodhart problem. The agents are winning at the metric and losing at life.

Remedy: Tie the reward to *composite* outcomes. For example, include HCU-like signals or direct terms for repair/relief/flexibility in the reward, and penalize trajectories that are non-compensable even if they hit the proxy. In short, make the game reward what actually matters for long-term balance.

18.6.3 One-agent fairness in a multi-agent world

Symptom: A single agent’s ledger looks wonderful by the end—but everyone else’s menu has collapsed. (One hero, many victims.)

Why it happens: There’s no coupling between agents’ choice sets. The “steak” runs out for everyone else, but the model didn’t remove steak from the hero’s menu, so one agent hogged all the benefit while others got wrecked.

Diagnostic: Track population-coupled admissibility. In a proper model, one agent’s admissible set $A_i(t)$ should *depend* (in part) on the state of others and shared resources. If you treat every agent in isolation, you miss this.

Remedy: Implement shared-resource constraints, congestion costs, and group-level compensatory options. In other words, when one agent’s behavior would ruin others’ ledgers, the simulation should introduce something (or restrict something) so that doesn’t spiral out of control.

18.6.4 Endgame without tilt

Symptom: As simulated death or closure nears, agents keep pursuing the same mix of actions as before—indulgent or short-sighted policies—without any surge of reparative or meaningful choices. The endgame is as chaotic as the middle. Why it happens: The shadow price is constant or non-existent; the horizon H_t never raises the cost of uncompensable moves. In the model, there’s no extra “gravity” pulling choices toward closure as $t \rightarrow T$.

Diagnostic: Estimate the $\Phi \times H^{-1}$ interaction (i.e. how compensability factors change as horizon shortens). If it’s ~ 0 , your model lacks “endgame physics.” The agents don’t act any differently when time is almost up.

Remedy: Make the admissibility threshold and option weights sensitive to H^{-1} . In practice, that means increasing the penalties for irreparable moves (or boosting repair moves) as the horizon shrinks. Late in life (or a mission), only moves that won’t leave permanent debt should pass the filter.

18.6.5 Dreamless agents (no offline counterweights)

Symptom: Agent policies get brittle; they cycle in local ruts and accumulate “debt” because they never step back. Essentially, they double-down on failing strategies. Why it happens: There is no off-policy recombination phase (no sleep/dream analogue) to reopen closed channels. The agents never get that creative, low-risk practice run to rethink things, so they stay stuck.

Diagnostic: Try inserting short “offline” epochs where agents replay high-debt scenarios with low real-world cost (like a dream sequence) and see if their menu broadens afterward. If adding that phase suddenly improves everything, then the lack of it was the issue.

Remedy: Give the agents an explicit REM-like pass. In simulation terms, add a phase of simulated experience focused on repair/relief trajectories (without heavy penalties) and then feed the outcomes back in. This can be as simple as a reset of certain internal states combined with simulated scenarios, or as complex as training a separate dream model. The key is to allow safe experimentation that can influence real policy.

18.6.6 Cosmetic composites

Symptom: The model claims to use a *composite affect measure* (like HCl/HCU), but in practice all its channels are highly collinear or just trivial transformations of the main reward. In other words, it’s pretending to be multi-dimensional while effectively one-dimensional.

Why it happens: The designers built a “composite” by summing a bunch of proxies that all track the same latent variable (e.g. heart rate, blood pressure, and breath rate all track stress). So it’s HCl in name only; there’s no real breadth.

Diagnostic: Run a factor analysis or check incremental variance contributions. Demand that each channel in the composite provides some non-redundant signal. Also look for temporal dissociations (e.g. a physiological relief that precedes subjective relief, indicating they aren’t identical signals). If all channels move in lockstep, your composite is cosmetic.

Remedy: Use heterogeneous inputs for your affect index—self-report, physiology, neural signals, behavioral cues, dream content, etc.—and *preregister* criteria for each channel’s contribution. For example, require that removing any one channel changes the outcomes (or at least, test that). Ensure that your composite really is one, not a disguised scalar.

18.6.7 Optional-stopping exploits

Symptom: Simulation outcomes look “balanced” or as expected only because results were harvested at a convenient time, or runs were stopped once a target pattern appeared. In other words, the experimenter peeked and stopped when things looked good.

Why it happens: Analyst degrees of freedom. The researchers might not even realize they did it, but they effectively cherry-picked the endpoint.

Diagnostic: Preregister your stopping rules. Better yet, run everything with fixed horizon lengths or under blinded pipelines where you don't get to decide when to stop. If the effect disappears when you remove human fudging of endpoints, it wasn't real.

Remedy: Lock your analysis code and plans beforehand. Use adversarial collaborators or bots that enforce strict stopping rules. Simulate the same scenario with a variety of predetermined end times. Basically, take human "wiggle room" out of the equation when evaluating fairness outcomes.

18.6.8 Adversarial fit parity

Symptom: A rival theory or model (say, a plain reinforcement learner with some risk aversion, or a predictive-coding variant) matches all the key LoF signatures in your simulation. In other words, your LoF model doesn't outperform a simpler non-LoF model on the measured patterns.

Why it happens: The tasks in the simulation were too easy or not discriminative. Multiple mechanisms can explain the data because the challenges never forced the uniquely LoF behavior to surface. It's like having two different engines that both idle fine because you never hit the gas.

Diagnostic: Introduce stress tests that specifically target LoF's unique predictions: (i) scenarios that require horizon-contingent menu shrinkage (only LoF would naturally do that), (ii) situations where an action has equal immediate utility as another but differs in long-term compensability (LoF would favor the one with repair potential), (iii) group scenarios where only a population-coupled mechanism would prevent a fairness collapse. If your rival model handles all those too, then indeed it's matching LoF in function.

Remedy: Design adversarial benchmarks that pit LoF against plausible alternatives. And when you run them, compare models with proper statistical criteria (WAIC for predictive fit (with appropriate held-out evaluation), or BIC as a simpler complexity-penalized fit criterion, ablation tests, etc.). The goal is to identify where LoF *actually outperforms* or where it fails.

18.6.9 Non-identifiability of the ledger

Symptom: You can "prove" neutrality in your sim only by assuming hidden variables or parameters that *guarantee* it. In other words, your ledger is defined in a circular way (you bake neutrality into the model's guts).

Why it happens: The ledger $L(T)$ is defined implicitly from the same data it's supposed to predict. Perhaps you inferred people's ledgers using the assumption that ledgers end at zero. That's a self-fulfilling prophecy, not evidence.

Diagnostic: Separate *measurement* from *mechanism*. For example, derive $L(T)$ purely from pre-registered HCI inputs (keep that process blinded to outcomes), *then* test whether LoF dynamics hold on that fixed series. If you have to assume what you're trying to prove, you're in trouble.

Remedy: Use two-stage pipelines, cross-validation, and external validity checks. For instance, calibrate your HCI and ledger on one dataset (maybe daily life data), then test LoF predictions on another (say, hospice patients) without tweaking the measure. The idea is to not let the fox guard the henhouse: the ledger measure should stand on its own, and then you see if LoF holds.

18.6.10 Fairness leakage at scale

Symptom: Your small simulations (with a handful of agents) behave well, but when you scale up to many agents, you start seeing cascading inequities, permanent underclasses, or widespread griefing that the model doesn't correct. The fairness you thought you had "leaks" away at population scale.

Why it happens: Constraints that work locally aren't propagating globally. There's no higher-level governance layer to handle large-scale interactions or resource constraints. Essentially, the model's LoF implementation might handle 1-on-1 fairness but not many-to-many fairness.

Diagnostic: Stress-test with scenarios like cohort shocks (suddenly 50% of agents get hit with something), resource scarcity, or coordinated strategies by subgroups. Plot the variance and tail risk of ledgers over time as the population grows. If the tails keep growing or a subset keeps getting worse off, your fairness isn't holding.

Remedy: Add a *meso-level* or institutional layer to the constraints: norms induction, restorative justice affordances, "cooldown" periods that scale with aggregate harm, institutional repair channels (like global events that encourage reconciliation). In short, design something akin to social institutions into the model to catch what individual-level QoS cannot.

18.6.11 The “nice dataset” trap

Symptom: All the LoF effects appear in clean, toy environments but vanish when you use messier data or include human participants. Everything worked in the sandbox, nothing works in the wild.

Why it happens: Overfitting to idealized conditions, ignoring real-world complications like missing data, nonstationary dynamics, and agents who adapt to the experiment. Essentially, you proved LoF under best-case assumptions, not in reality.

Diagnostic: Port the same model or experiment to a human-in-the-loop task or a noisy real dataset. Track what breaks. Also introduce known challenges in simulation: missing data points, environment shifts, agents that learn the experimenter's patterns. If performance degrades significantly, you've been in a "nice dataset" bubble.

Remedy: Build robustness checks into your pipeline. Inject noise, simulate domain shifts, remove chunks of data, add agents that deliberately try to mess up the scenario. Have clear criteria for what counts as acceptable degradation. And be prepared to report that robustness (or lack thereof) honestly.

A preregistration checklist for LoF simulations: To avoid many of the above pitfalls, any planned simulation of LoF should commit to some key design choices before running:

- Mechanism vs. makeup: No resets that directly zero the ledger (mechanisms must do the work, not manual resets).
- Composite affect: Use a multi-channel HCI and test that each channel adds value (no single-metric monoculture).
- Horizon logic: Implement an explicit λ_t (shadow price) that increases as H_t^{-1} does (i.e. as the horizon shortens).
- Population coupling: Ensure agents' menus can respond to other agents' actions and shared resource levels.
- Offline phase: Include a dream/REM-like offline processing phase and look for its effects on subsequent behavior.
- Rival baselines: Include strong comparison models (no-fairness models, or alternative fairness mechanisms) and choose tasks that can tell them apart.
- Blinding and stopping: Use locked analysis code, fixed or strictly rule-based horizons, and declare endpoints/timelines in advance.
- Robustness: Plan noise injections, domain shifts, and scale-up tests with predetermined pass/fail criteria.

Reporting nulls and near-misses (what honesty looks like):

- If neutrality *only* holds when you include resets or external fixes, label that clearly as "neutrality-by fiat," not as partial support for LoF.

- If rival models tie with your LoF model, say so, and specify what new tests or data could differentiate them next time.
- If scaling up breaks fairness, document exactly where the coupling failed and propose adding the necessary governance-level constraint (don't just gloss over it).
- If your composite affect measure collapses to effectively one channel, admit that and either rebuild the measure or downgrade your claims accordingly.

Takeaway: A good LoF simulation doesn't force fairness; it forbids unfair trajectories by shaping what options persist—especially as horizons shrink and populations interact. When we observe any of the failure patterns above, it means we've either mistaken patches for physics or built worlds too weak to tell a true law from a convenient tendency. Keeping these traps in mind keeps our science honest—and it prepares us for the ethical questions that follow in the next part of the book.

18.6.12 Where we go next:

Moving on to Part IX, we leave empirical and theoretical questions and confront the ethical horizon. In the next part, Chapter 19 (“What This Never Justifies”) will make absolutely clear that, true or false, LoF never licenses harm or complacency. We will enumerate the inviolable ethical principles that guide caregivers, researchers, and society—ensuring that human dignity and compassion remain front and center as we conclude our exploration of the Law of Fairness.

Part IX — Ethics and Human Dignity

Scientific ambition has led us through complex theories, equations, and bold experiments. But at some point in this journey, a simple question arises: *Even if we can prove a Law of Fairness, should we pursue it at any cost?* Imagine a research team considering an end-of-life trial to measure every flicker of pain and relief in a dying patient. The data might be invaluable—but should they ever withhold comfort for the sake of science? Questions like these bring us face to face with ethics and human dignity. This part steps back from the technical details to confront the moral boundaries of our exploration. After all, unlike a physics experiment with mindless particles, here we are measuring conscious experiences—real human pain and joy—which means the stakes are profoundly personal.

We've posited that each conscious life must end with a neutral ledger of felt experience. Under LoF, the system must balance out in the end, but that is a constraint of nature—not a license to cause harm. No scientific insight, however groundbreaking, can justify violating a person's dignity or well-being in the name of data. History offers painful lessons of what can happen when curiosity trumps ethics: from infamous studies that withheld treatment from suffering patients to extreme experiments that traumatized participants, the pursuit of knowledge has sometimes strayed into cruelty. We are determined not to repeat those mistakes. That is why we insist on treating the Law of Fairness as *descriptive* science, not a moral prescription. It predicts a balance of joy and suffering by life's end, but it carries no mandate to "enforce" that balance or treat people as mere data points. If someone is in pain, we are duty-bound to help relieve it, not stand aside waiting for some theoretical equilibrium. In short, human beings remain at the center of the story—not as entries in a ledger, but as individuals whose comfort and autonomy matter above all else.

It's also crucial to remember that LoF applies to each unified stream of consciousness individually, not to humanity as a whole. You cannot "balance" one person's pain with another person's pleasure—each life's ledger stands on its own. This means we can never justify sacrificing one person's well-being on the theory that it might boost someone else's happiness or serve some greater equilibrium. Fairness, in this framework, isn't about trading off between people; it's about the natural balancing that may occur *within* one life, on its own terms. Every individual's story matters, and no one's suffering is a legitimate currency for someone else's relief.

Research Notes: Ethical Safeguards. Rigor and ethics go hand in hand. We set strict end-of-life neutrality gates to declare a life's ledger 'neutral' only with compelling evidence (final hedonic average within a preregistered standardized tolerance band

around zero, last-phase slope within a preregistered near-zero bound, and emotional variance not exceeding a preregistered proportion of baseline phase variance). Any comparisons across different groups or cultures follow a measurement invariance ladder (Configural → Metric → Scalar) to ensure we’re measuring the same construct; if a step fails (e.g., metrics differ by group), we confine conclusions to within-person patterns. All studies were preregistered, and each study named a single out-of-sample metric (such as WAIC or log-loss) ahead of time to evaluate model fit on new data. These safeguards aren’t just procedural—they uphold an ethical commitment to honest, unbiased science. By design, we avoid bending rules or overstating results in ways that could mislead lives or policy.

In the chapters that follow, we address two sides of the ethics equation. First, we draw clear lines around what the Law of Fairness never justifies when applied to real human situations, establishing inviolable guardrails for research and practice. Next, we consider how the very idea of an inevitable balance of experience might affect one’s personal outlook—exploring whether LoF leaves room for hope, freedom of choice, and meaning in daily life. By confronting both the practical ethics and the existential implications, Part IX guards against misuse of the theory and nihilistic misreadings. By the end of this part, you’ll see how scientific integrity and human dignity are woven together in our approach, ensuring that the pursuit of knowledge never loses sight of compassion and respect.

What this Part will do for you:

- A clear understanding of why the Law of Fairness is *descriptive* science, never a moral license to inflict pain or withhold care.
- The ethical guardrails guiding any LoF research or application, including the principle that a subject’s comfort and dignity always outweigh the value of additional data.
- Insight into how LoF’s guarantee of balance can be respected without undermining free will or the personal significance of joy and suffering.
- A perspective on integrating the idea of an “equalized” life ledger with everyday hope and decision-making — while avoiding fatalism or the sense that life’s experiences don’t matter.

Chapters in this Part:

- **Chapter 19 — What This Never Justifies** - explores the non-negotiable ethical limits of this theory, making sure no one misuses LoF to excuse cruelty, neglect, or indifference to suffering.
- **Chapter 20 — Hope, Freedom, and Daily Life** - examines how a guaranteed balance of experiences can coexist with genuine hope, personal freedom, and the search for meaning—ensuring that LoF enriches rather than diminishes everyday human life.

Where we go next:

With these principles established, we now turn to Chapter 19, which asks a critical question: What does the Law of Fairness never justify? In the next chapter, we examine the limits of fairness-driven reasoning, beginning with an uncompromising look at why no pursuit of balance can ever warrant the violation of individual dignity (19.1).

Chapter 19 — What This Never Justifies

Picture a hospice ward in the quiet hours of the night. A patient near the end of life is moaning in pain, her consciousness flickering as monitors record every heartbeat and breath. A researcher stands by, watching a real-time graph of her Hedonic Composite Index. The data could reveal whether her final moments tilt positive or negative — a crucial piece of evidence for the Law of Fairness. Now the patient pleads for more relief from the pain. The researcher faces an impossible choice: increase the morphine dose and ease her suffering, or hold off to avoid “spoiling” the data. It’s a ghastly scenario to imagine, and it captures the core question of this chapter: *Does the pursuit of knowledge ever justify withholding compassion?*

The answer is an emphatic no. This chapter lays out the non-negotiable ethical limits of our theory, starting with a simple guiding principle: “Relief is a systems variable; comfort and dignity override data collection.” In practice, that means no finding, theory, or measurement goal is ever more important than a person’s well-being. If a patient is suffering, we provide relief — full stop. We would never keep someone in pain just to see if a last-moment surge of happiness balances their ledger. The very idea violates basic human decency and the spirit of *fairness* itself. LoF would predict that all pain will eventually be compensated, but it does not give us permission to inflict pain or prolong misery to satisfy our curiosity.

Crucially, the Law of Fairness is a constraint of nature, not a moral blueprint or cosmic purpose. It doesn’t say that “suffering should happen” or that anyone ought to endure hardship for the sake of balance; it only hypothesizes that in the end, the ledger of felt experience *will* net to zero. There is no cosmic referee demanding that a happy person must be punished or a sad person must be rewarded before they die. To be clear: if someone’s life has been filled with joy, we have no business imposing pain on them in some misguided effort to “even things out.” And if another person’s life has been filled with sorrow, we can’t just assume the universe will grant them happiness eventually—we have to take an active role in easing their burden. Any balancing that occurs must emerge naturally or through compassionate action, not through any cruel or calculated scheme. In earlier chapters, we introduced a hypothetical “Queue System” mechanism that might drive a balance internally over time. But even if such a mechanism exists, it operates within the person’s own mind and body. We humans have no mandate (nor any rightful desire) to play enforcer of the cosmic ledger. Our job is to care, to observe, to measure carefully, and to never lose sight of empathy.

Decades of hard-won lessons in research ethics (often born from regrettable abuses in science) reinforce that people can never be treated as mere means to an end. However

valuable the data might be, the individuals providing that data are human beings with rights and dignity. All standard ethical safeguards – independent review, informed consent, the right to withdraw, and the duty to minimize harm – apply fully here. In designing any study around LoF, we choose methods that respect these boundaries. For example, researchers might conduct *observational* studies that track a patient's emotional state while they receive all appropriate care, rather than experimental studies that withhold relief. Our tools could include pain diaries, sensor bracelets, or noninvasive brain scans – approaches that gather insights without ever crossing a patient's comfort threshold. In short, we align our research with the principle that the path to discovery must be as humane as the goal is profound.

The phrase “relief is a systems variable” underscores that giving comfort isn’t an interference in what we’re measuring; it’s a valid input to that system. Administering a painkiller or offering emotional support simply adds a positive entry to the person’s ledger – helping to fulfill LoF’s balance in an ethical way. We incorporate those acts of kindness into our analysis rather than avoid them. There is never a conflict between doing the right thing for a person and collecting useful information; if a conflict seems to arise, it means we’re asking the wrong research question.

Finally, it’s worth dispelling any notion of fatalism or indifference. Some might worry that if life naturally balances pleasure and pain, one could become callous – thinking, for example, “This person’s suffering will be offset eventually, so maybe it’s okay if I don’t help.” That mindset is categorically wrong. LoF is not a karmic justice system or an excuse to be passive in the face of pain. Even if a drowning person *will* eventually breathe again in the long run, you still throw them a lifeline now. Fairness in the LoF sense is never a justification for inaction or cruelty. On the contrary, if the law holds, it highlights how precious relief and joy are when they occur – they are the forces that ultimately balance out darkness. It would be the bitterest irony to seek proof of a “fair” universe by behaving unfairly or inhumanely. This chapter cements the point that no matter what results we chase, our methods must remain compassionate, respectful, and above reproach.

What you’ll get from this Chapter:

- A clear statement of ethical priorities: why no data or theory ever justifies causing or prolonging someone’s suffering.
- The reasoning behind treating relief and comfort as integral parts of the system (rather than “interference”), so that helping a subject isn’t just allowed but encouraged in our research design.

- Concrete scenarios illustrating how we handle the tension between scientific curiosity and moral responsibility – always resolving it in favor of the person’s well-being.
- A refutation of any “ends justify the means” thinking with regard to LoF, including why the law’s existence would never excuse cruelty, neglect, or withholding care.
- Reinforcement that LoF is a descriptive claim about nature’s balance, not a green light to inflict pain or a cosmic mandate that anyone must suffer.

Subsections in this Chapter:

- **19.1 No License to Ignore Pain** - This subsection makes clear that belief in the Law of Fairness is never an excuse to neglect or delay relieving someone’s suffering. No matter what balancing might occur in the long run, caregivers must address present pain. In other words, *immediate compassion is non-negotiable* – one cannot justify inaction by assuming “it will all balance out” later.
- **19.2 Duties of Caregivers and Researchers** - This part lays out the responsibilities of those caring for others or studying them. It emphasizes providing comfort promptly, obtaining informed consent for any intervention or study, and respecting privacy. Crucially, if a procedure or experiment is causing distress without clear benefit, it must be stopped or redesigned. Observing suffering for data’s sake is never justified – caring for the person comes first, always.
- **19.3 Justice Aimed at Restoration** - Here we consider implications for justice systems. If something like LoF operates, the goal of justice should be to *restore balance* and rehabilitate rather than to exact retribution. Punishment for its own sake finds no support in a balance-oriented view. Instead, justice would focus on repairing harm and helping wrongdoers make amends, all while making decisions cautiously and based on evidence. This approach favors healing over hurting, aiming to re-integrate individuals rather than simply inflict penalties.
- **19.4 Research Notes: Non-Sentience in Simulation** - This research-focused subsection insists that we must avoid causing suffering in any experiments meant to probe LoF. It suggests using non-sentient agents (simulations, computer models) or humane, retrospective data when exploring the theory. In short, researchers should test hypotheses about balancing *without harming any conscious being*. If a study *might* create pain or distress in a sentient subject, alternative methods should be sought to safeguard welfare.
- **19.5 Communication Ethics** - Even how we talk about the Law of Fairness carries ethical weight. This subsection cautions against using simplistic or deterministic slogans (e.g. glibly saying “everything balances out” to someone in pain). It advises honesty about uncertainty and avoiding euphemisms that downplay suffering. Communicators should never offer false hope or promise comfort that

isn't supported by evidence. The takeaway is to speak about LoF with humility and compassion, never in a way that could trivialize someone's struggles or mislead them.

- **19.6 Hard Lines We Will Not Cross** - The final subsection draws firm ethical boundaries. It reminds us that providing relief and preserving dignity override any data-gathering or theoretical goals. Certain practices are absolutely off-limits: one must never withhold care deceitfully (e.g. pretending one can't help just to "see what happens"), never perform invasive or distressing procedures without consent, never coerce participation or behaviors, and never attempt to force a balance in someone's life. In essence, no matter what LoF suggests, we do not cross these moral lines – people's well-being comes first.

Where we go next:

We begin immediately with 19.1, which makes clear that *no* hypothesis of eventual balance can justify ignoring someone's present pain. First, we assert in plain terms why LoF, if true, only strengthens our moral duty to relieve suffering now.

19.1 No License to Ignore Pain

Claim in plain speech: Even if the Law of Fairness (LoF) is true as a constraint on life-long experience, it never diminishes the moral urgency of pain in the present. LoF is a scientific hypothesis about aggregate hedonic accounting; it is not a voucher to delay care, rationalize harm, or sermonize that “it will all balance out.” If anything, LoF strengthens the case for immediate relief and repair: by the Queue System’s logic, reducing current load and expanding compensable options is precisely how neutrality becomes reachable.

19.1.1 What LoF says here—and what it doesn’t

- Descriptive, not prescriptive. LoF describes a guardrail on how feelings can accumulate over an intact stream of consciousness. It does not prescribe that anyone should suffer now so that some later “balance” occurs.
- Menus, not mandates. QS prunes and weights options; it doesn’t command outcomes. Agency remains inside the admissible set. Helping someone in pain widens that set (sleep, analgesia, support, reconciliation), increasing the feasibility of compensation.
- Local experience is decisive. The only feelings that exist are the ones happening *somewhere to someone right now*. LoF cannot be invoked to discount those feelings without committing a category error.

19.1.2 Five harmful misuses (and the scientific/ethical correction)

1. Cosmic offsetting (“Suffer now, it evens out later”).
 - *Why it’s wrong scientifically:* You do not control the future path or the horizon; compensability is uncertain and state-dependent.
 - *Why it’s wrong ethically:* It treats the current person as a means to a hypothetical end.
 - *Correction:* Relief-first—reduce load now, then enable repair.
2. Moral discounting (“Your pain counts less because of the long run”).
 - *Wrong scientifically:* The ledger is an *integral*; every unit experienced now is fully counted.
 - *Wrong ethically:* Dignity attaches to present persons, not abstractions.
 - *Correction:* Treat present distress as fully real and urgent.

3. Fatalism (“My choices don’t matter if neutrality is guaranteed”).
 - *Wrong scientifically:* Choices alter menus—for you and for others (menu coupling).
 - *Wrong ethically:* Responsibility tracks foreseeable effects on others’ options.
 - *Correction:* Emphasize option-widening actions and reversibility.
4. Withholding care (“Let the system balance them”).
 - *Wrong scientifically:* Non-compensable burdens can collapse menus, making balance *harder*.
 - *Wrong ethically:* Neglect is harm.
 - *Correction:* Provide care early; prevent spirals that shrink future options.
5. Narrative spiritualizing (“Pain proves a cosmic plan”).
 - *Wrong scientifically:* LoF posits a constraint, not a purpose.
 - *Wrong ethically:* Risks silencing people who need help now.
 - *Correction:* Keep metaphysics out of triage; offer practical support.

19.1.3 Immediate-relief doctrine (the stance that fits LoF best)

Principle: When someone is suffering, the ethically and scientifically aligned response is to lower current load and raise future flexibility.

Three levers you can pull today:

- Load reduction: sleep, analgesia, anxiolysis; quieting environments; financial or logistical relief.
- Repair enablers: access to people and tools that close loops (mediators, clinicians, case managers, debt counselors).
- Flexibility/reversibility: cooling-off periods, second-chance policies, easy exits from commitments that became harmful.

Everyday examples.

- *Clinic:* A patient in severe pain is offered immediate symptom relief before complex diagnostic workups.

- *Workplace*: A struggling employee is granted a temporary workload reduction and a path to repair performance without stigma.
- *Family*: In conflict, parties pause, sleep, and reconvene with a mediator rather than escalate when horizons feel short.

19.1.4 Language that helps vs. language that harms

Say this:

- “Your pain matters right now. Let’s reduce the load first.”
- “We can widen options—here are three things we can try today.”
- “This framework is a research hypothesis. Help and safety come first.”

Avoid this:

- “It will even out.” / “Everything happens for a reason.”
- “Try to see the bigger picture.”
- “This is a test you’re meant to endure.”

If someone mentions self-harm or suicide:

- Acknowledge and validate: “I’m really glad you said that; I take it seriously.”
- Move to safety: share local crisis resources and stay with them (virtually or in person) until connected; remove means if physically present; involve trained professionals per crisis protocol.
- Do not debate cosmology, balance, or purpose. The only correct next step is care and connection.

(For online communities, always include region-appropriate helplines and escalation protocols. Use non-judgmental, non-sensational language; no details about methods.)

19.1.5 Designing systems that don’t ignore pain

Policy design heuristics (usable by schools, clinics, workplaces, platforms):

- Access first: Make relief services the *default path* (opt-out, not opt-in).
- Reversibility: Build “undo” and “cool-off” features; reversible choices keep future menus larger.

- Low-friction repair: Create standard, non-punitive channels for apology, restitution, and re-entry.
- Menu coupling audit: Before launching policies, ask “Whose options shrink?” and “How do we widen theirs?”
- Horizon-sensitive support: Increase outreach near predictable closures (exams, deadlines, layoffs, terminal care).

19.1.6 For clinicians, caregivers, and moderators: compact checklists

Clinician/Caregiver 6-point check:

1. Assess and treat pain/distress promptly (don’t gate relief on long lectures).
2. Offer at least two immediate, reversible options that lighten load today.
3. Activate social supports (warm handoffs, not just referrals).
4. Normalize repair (“It’s common to need to mend this; here’s how”).
5. Document safety; if any risk language appears, follow crisis protocol.
6. Schedule near-term follow-up to maintain widened menus.

Community moderator 6-point check:

1. Validate; avoid cosmological framing.
2. Share crisis resources; invite DM for a private check-in.
3. Enforce content rules that prohibit glamorizing self-harm.
4. Remove/flag harmful replies; nudge toward help-seeking.
5. Keep posts about LoF clearly labeled as theoretical.
6. Log and review incidents to improve protocols.

19.1.7 Research and ethics: no “balance later” designs

- Consent: Plain-language statements that participation will not withhold indicated care.
- Harm minimization: Stop rules when distress crosses thresholds; independent monitors.
- Preregistration: Hypotheses, outcomes, and analysis plans are set before data collection.

- Debriefing and resources: Every participant leaves with support options.
- Non-sentience by default in simulations: No affective states in code without committee review, external audit, and a clear scientific necessity (expanded in 19.4).

19.1.8 Special topics

Justice and LoF. “Evening out” is not punishment; it’s *restoration*. Systems aligned with LoF expand compensable pathways for those historically denied them (health access, education, clean air, fair process). See 19.3.

Public claims. Any public statement about LoF should carry two disclaimers:

1. “This is an evolving research framework; it may be wrong in part or whole.”
2. “No claim here replaces clinical care or crisis support.”

19.1.9 Where we go next:

Having established the primacy of relieving pain, we turn in 19.2 to the concrete responsibilities of those who care for others or conduct research. Next we will translate this “relief-first” ethos into formal duties and guidelines for caregivers, community moderators, and scientists.

19.2 Duties of Caregivers and Researchers

Thesis: If LoF is a descriptive constraint on life-long affect, then the normative duties for people who hold power over others' options are clear: relieve present load, widen future menus, and never trade on some hypothetical "balance later" to justify current harm. This section translates that stance into concrete obligations for clinicians, caregivers, community moderators, and researchers alike.

19.2.1 First principles (shared across roles)

All helping and research roles should abide by these foundational ethics:

- **Beneficence now.** Prioritize the immediate reduction of suffering (pain relief, de-escalation, ensuring safety) *before* indulging in education or theory. Urgent help comes first.
- **Non-maleficence.** Do not introduce new burdens – pain, shame, stigma, financial strain – unless they are strictly necessary, proportionate to a legitimate aim, and accompanied by mitigation. And even then, minimize and justify every harm.
- **Respect for autonomy.** Provide clear, plain-language choices. Avoid "nudges" that hide value judgments behind "the science." People must be free to choose or refuse interventions without coercion or deceit.
- **Justice.** Audit who gains options and who loses them under any plan ("menu-coupling" analysis). Actively seek out and correct inequities – don't assume a policy is neutral; check its differential impact.
- **Dignity.** Always speak to *persons*, not ledgers. In triage or crisis, do not spout cosmic theories or moralize about eventual neutrality. No one in pain should be made to feel like a datapoint or a cog in some grand balancing machine.

19.2.2 Duties of caregivers (clinicians, counselors, moderators, case workers)

B1. The Load-First Protocol – In any caregiving context, when someone is in distress:

- Assess and treat distress promptly. Don't gate relief behind paperwork or protracted evaluations. Address physical pain and emotional crisis *immediately* as a first step.
- Offer reversible, near-term relief options. Provide at least two immediate options the person can take now to lighten their load (e.g. a low dose vs. medium dose, a cooling-off period, a quiet room to rest) – and ensure these options are reversible so they don't feel trapped.

- Activate supports with warm handoffs. Don't just hand someone a brochure or referral. Connect them directly (with permission) to family, peer support, social worker, etc. – ensure they actually get support, not just an FYI.
- Create a repair path. If harm has occurred (an offense, mistake, conflict), establish a clear and *non-punitive* path to make amends or repair (e.g. a mediated apology, a restitution plan) with specifics (what must be done, with whom, by when). Emphasize that needing to repair is common and not a permanent stigma.
- Follow up soon. Shrinking horizons (someone in crisis, at end-of-life, etc.) need *frequent touchpoints*. Schedule a follow-up within 24–72 hours to keep their menu of options from collapsing again.

B2. Suicide and self-harm – Special handling for any mention or sign of self-harm:

- Take every mention seriously. Never dismiss or downplay expressions of suicidal ideation or self-harm. Validate the person's feelings and pivot immediately to a safety plan.
- Escalate to trained help. Follow your crisis policy: involve mental health professionals or crisis teams right away; do not try to "reason it out" with them using LoF or any theory. This is a time for emergency care, not discussion.
- Reduce access to means (if in your capacity). Where feasible and legal (e.g. if you're physically present and they consent or it's an emergency), remove or secure any means of self-harm. Stay connected with the person until you can transfer care to a professional or trusted support who will keep them safe.

B3. Communication standards – How caregivers talk about LoF or suffering:

- Do say (emphasize the now): "Your pain matters right now. Let's reduce the load and widen your options today."
- Don't say (no platitudes): "It'll even out in the end," or "This is happening for a reason," or "Just endure it for the lesson." Such phrases, even if well-intended, *invalidate* the person's present pain and can deepen despair.
- Document facts, not theory. In clinical charts or case notes, stick to observations, options offered, what the person decided, and next steps. Do not wax philosophical about LoF or neutrality in official records – theoretical talk belongs in research forums or education sessions, *not* in the midst of someone's crisis.

B4. Equity and menu-coupling checks – Before implementing any policy (a hospital discharge criterion, a community moderation rule, etc.), run an equity audit: ask yourself and your team:

1. Whose choice set might shrink because of this rule? (E.g. Does it burden people with less money, less time, or other constraints more than others?)
2. What compensatory supports are in place for those folks? (If a new policy limits something for the most vulnerable, what are we giving them to offset that loss?)
3. Are the burdens falling on those with the least horizon? (E.g. Does it hit the sickest patients, or the poorest users, the hardest?) If yes, rethink.

Adjust the policy until it widens options (or at least minimizes harm) for the most constrained parties. The goal is a policy that doesn't just serve the average, but specifically *does no further harm* to those whose menus are already tight.

19.2.3 Duties of researchers (human subjects research, data practices, and simulation work)

C1. Consent that *actually* informs – Every study consent form or script must truly educate and protect the participant:

- Plain language: Use clear wording like: "*This study will not delay or replace any treatment you would otherwise receive. You can stop at any time without penalty.*" Eliminate jargon and make sure a layperson understands their rights.
- Scope boundaries: Be upfront about what the study is and isn't. E.g.: "*This project studies feelings about X; it is not a therapy and will not provide treatment.*" Participants should know this is observational or experimental, not an intervention for their benefit.
- Risk disclosure: List any foreseeable risks (fatigue, emotional triggers, privacy risks, etc.) and the steps you've taken to mitigate them (breaks allowed, option to skip questions, clinician on call, etc.). Do *not* hide discomfort or surveillance under technical language.
- No metaphysics in consent: If LoF or related theory is mentioned, frame it clearly as a hypothesis you are testing – *not* a proven fact and certainly not a promise of any personal outcome. Participants should not be misled into thinking "this study will prove life is fair" or any such grand claim.

C2. Harm minimization in design – Plan studies with the default that participant welfare is paramount:

- Relief-first rule: If a participant’s distress goes beyond a minimal threshold, *pause the study and help them*. For example, if someone becomes visibly upset during a survey, you stop and offer support. Only continue if an independent evaluator agrees it’s safe and the participant still consents.
- Horizon-sensitive scheduling: Do not schedule demanding sessions at times when people’s personal “horizons” are naturally strained (end of a long shift, late at night, right after a major loss, etc.). Build in flexibility for rescheduling so that participants aren’t pushed when they’re vulnerable.
- Minimal deception: Deceptive methods (if truly necessary) must be *extremely limited, fully IRB-approved*, and always followed by a thorough debrief and an offer for participants to withdraw their data afterward. No deception that could cause lasting mistrust or trauma is allowable under LoF ethics.

C3. Analysis integrity – Ensuring our scientific practices are rigorous and fair:

- Pre-registered plans: Formally preregister your hypotheses, outcome measures, exclusion criteria, and analysis plan before collecting data. This prevents cherry-picking “balanced” outcomes after the fact.
- Blinding where possible: Use blinded analysis pipelines so that, for example, analysts don’t know which participants were in which group when scoring subjective outcomes. This reduces bias.
- Negative controls and adversarial comparisons: Include checks for things LoF *should not predict* (negative controls) and compare against credible rival models. Define these rivals in advance. This way, you’re testing LoF fairly rather than bending results to fit it.
- Missing-data plans: Have a strategy so that data from vulnerable subgroups *is not simply dropped*. (E.g., if marginalized participants tend to drop out more, your analysis should address that rather than silently excluding them and declaring “balance” based only on the remainder.)

C4. Data dignity – How we collect and handle data about feelings and lives:

- Minimize collection: Gather only the data that is essential for the hypothesis. No “collect everything just in case” hoarding. If you don’t absolutely need a piece of personal info, don’t collect it.
- Secure and limit access: Use local encryption and strict role-based access controls for any stored data. Default to *short retention* of raw data unless

participants explicitly opt in for long-term archiving of their information. Data about someone's inner life should not live indefinitely on a server without good reason and consent.

- Participant agency with data: Give participants the right to view their data, annotate it (e.g. "this doesn't represent how I felt"), or request portions be deleted if they feel it misrepresents them. They are not just data points; they should have a say in how their information is used.
- Community review: For studies dealing with culturally sensitive content or vulnerable populations, involve representatives from those communities in reviewing the protocol. They might catch ethical issues or misinterpretations that researchers miss.

C5. Simulation guardrails (*preview of 19.4*) – If you are doing computational or AI simulations to test LoF-related ideas, enforce these rules:

- Non-sentience by default: Assume no simulated agent should have any capacity to *feel*. Do not create agents with architectures that even *might* generate conscious experience (no matter how rudimentary) unless you undergo explicit ethics review and external audit, and have "kill-switches" ready. In normal work, keep agents clearly below any sentience threshold.
- No "neutrality by fiat": Do not hard-code outcomes to force a balance. (E.g., don't just program your simulation to always end up neutral.) Let any balance or compensation effects emerge (or not) from transparent constraints. Forcing a result defeats the purpose of testing LoF and is misleading.
- No public exposure of potentially sentient sims: Never deploy a simulation to the public (or anywhere uncontrolled) if it *could plausibly be sentient*. In LoF research, we have zero tolerance for accidentally creating digital suffering. Keep experimental AI contained and clearly non-sentient.

19.2.4 Role-specific checklists (*tear-out cards for quick reference*)

Clinician/Caregiver 6-Check:

1. Load reduced?
2. Two reversible options offered?
3. Warm handoff made to support?
4. Repair path set (if needed)?

5. Safety plan in place (if any risk language used)?
6. Near-term follow-up booked?

Principal Investigator/Research Lead 6-Check:

1. Consent truly plain and understood?
2. Distress stop-rules defined *and active*?
3. Preregistration posted (with adversarial models)?
4. Blinds/controls implemented where possible?
5. Data minimized and participant rights (view/delete) enabled?
6. Independent monitor appointed for oversight?

Community Moderator 6-Check:

1. Validate the person; move off theory when someone's in crisis.
2. Provide local help links or resources promptly.
3. Remove or hide any glamorizing self-harm content immediately.
4. Offer to DM and provide a warm handoff to professional help if appropriate.
5. Log incidents for review and learning.
6. Reiterate: "LoF is a research hypothesis, *not* clinical guidance or life advice."

19.2.5 Sample language you can reuse:

- *Clinical intake addendum (for healthcare settings)*: "Our priority is to reduce your distress today and expand your options for tomorrow. Any discussion of theories comes second to your care." (A gentle reminder that patient comfort comes first, theory second.)
- *Study consent paragraph (for research forms)*: "This is a research study about how people feel and make decisions. Taking part will not delay or replace any treatment you would otherwise receive. If you feel worse at any point, we will pause and connect you with help. You can stop at any time."
- *Community banner (for forums discussing heavy topics)*: "We welcome discussion of ideas, but if you're struggling or thinking about self-harm, please reach out — help is available. Theory never replaces care, and you are not alone."

19.2.6 Accountability and learning:

- After-action reviews for any adverse events, with representation from the affected parties. (If something goes wrong – a participant had a breakdown, a harmful post slipped through – convene a review that includes people like that participant or community members to honestly assess and improve.)
- Public summaries of what went wrong and what's changing. De-identify personal data, but be transparent about mistakes and how you're addressing them. This builds trust and a culture of improvement.
- Reciprocity: Offer participants and communities something back beyond "thanks." This could be aggregate results of the study, training sessions, resource lists, or skills workshops that benefit them. If people give us their time and data (and certainly if they take any risk), we owe them real value in return.

19.2.7 Where we go next:

In 19.3, we widen the scope from individual duties to society. Next we ask how justice itself would operate under LoF's constraint: instead of retribution, a fairness-aligned justice system would pursue restoration, aiming to repair harm done and keep everyone's life "ledgers" as balanced as possible.

19.3 Justice Aimed at Restoration

Thesis: If LoF is about conserving lifetime affect at a neutral equilibrium, then justice (in society, institutions, communities) should emphasize repair over retaliation. Just systems would cap harm, prefer restoration, and widen future choice-sets for everyone touched by an offense – those harmed, those responsible, and the community whose “menus” were narrowed by the event.

19.3.1 What “restoration” means here

When we talk about restorative justice in the context of LoF, we mean attending to *all parties* in a structured way:

- Primary restoration (to the harmed): Immediately reduce the harmed person’s load (ensure safety, provide care or compensation), restore specific lost options (give back time, access, status if possible), and create credible paths to longer-term well-being. The harmed individual should tangibly regain what was taken from them, as much as possible.
- Secondary restoration (to the responsible party): The person who caused harm must acknowledge it and accept proportionate consequences, *and* be given structured routes to make amends and rehabilitate – e.g. making restitution, building skills or understanding (competence-building), and a supervised plan for re-entry into trust. We aim not to destroy the person, but to help them reintegrate after genuine accountability.
- Tertiary restoration (to the community): The community must have its trust and predictability rebuilt. This means clear norms are reinforced, processes are transparent, and bystanders aren’t left with lingering fear. We also want to prevent the incident from shrinking others’ options (e.g. people withdrawing from public life out of fear). Community-level restoration might include public explanations, reforms, or support channels for those indirectly affected.

19.3.2 Guardrails that follow from LoF

A justice system aligned with LoF would set these policies:

- No “balance” via new uncompensable pain. You don’t achieve fairness by inflicting additional harm that cannot be repaired. Punishments that create *new* permanent wounds violate the spirit of LoF and sabotage neutral closure for multiple streams of experience at once. In short: adding trauma to trauma does not yield balance.

- Proportionality with reversibility. Favor responses that are reversible or can be lifted as the person makes amends. For example, instead of an irredeemable life sentence (except when absolutely necessary for safety), use graduated sanctions: probation that can end upon milestones, temporary restrictions that are reviewed periodically. Options should reopen as repair happens.
- Menu-widening bias. A just outcome should leave the harmed party with more viable options than they had immediately after the offense (their world should open up again), and it should provide the responsible party with specific opportunities to earn their way back into good standing. If an outcome only leaves everyone worse off and with fewer paths forward, it's likely not aligned with LoF.
- Horizon sensitivity. If someone has a short horizon (e.g. precarious health, nearing end-of-life, extreme poverty), any sanction will hit them harder. Justice must account for that. Identical penalties can have vastly unequal effects depending on the person's horizon and capacity. Thus, extra protection or adjusted consequences are warranted for those already severely constrained – otherwise we inadvertently crush those least able to bear it.

19.3.3 A practical sequence (*for courts, schools, workplaces, platforms when addressing wrongdoing*):

- Stabilize and verify harm. First, ensure everyone is *safe* and immediate needs are met. If someone was harmed, get them care (medical, psychological, etc.). Simultaneously, gather the facts of what happened. Importantly, avoid moralizing or assigning blame in this *acute* phase – focus on facts and safety.
- Name specific losses as “ledger items.” Clearly identify what was taken or damaged by the offense: time, money, health, trust, opportunities, reputation, etc. Each of these becomes an item on a “ledger” of harm that we will seek to repair. (This makes the harm concrete, not abstract “evil” or “badness,” but tangible deficits to address.)
- Co-design a repair plan with the harmed party. Work *with* the person (or people) who were harmed to outline how to make things as right as possible. This might include a restitution schedule (paying back stolen money or covering medical bills by a certain timeline), service that benefits the harmed or community, a public or private apology, and guaranteed follow-ups to track progress. The harmed individual’s needs and ideas should inform this plan.
- Build a competence track for the responsible party. In addition to making amends, the one who caused harm should have a structured program to improve

themselves and prevent repeat harm. This could involve training (e.g. anger management, bias training), mentorship, or monitored duties. Link their progress to staged restoration of privileges/trust. For example, after completing certain steps and demonstrating change, they gradually regain responsibilities or rights they lost.

- Create community scaffolds. Set up supports for the wider community: mediated forums for open dialogue if appropriate, clear norms re-communicated to all, bystander support channels (so others can report if they feel unsafe), and transparent reporting of how the case is being handled (to rebuild trust). The community should see that there is a fair process and that their environment will be safe.
- Review and adapt periodically. At set intervals, gather all parties (or their representatives) and evaluate how the plan is going. Have milestones been met? If repair is ahead of schedule, can we relieve some restrictions early (as a positive reinforcement)? If there's non-compliance or new issues, do we need to adjust or escalate? The idea is continuous feedback: reward genuine progress with more freedom, and only escalate consequences when someone refuses to comply despite support and chances.

19.3.4 Tools that operationalize restoration

To put the above principles into practice, consider using formal tools:

- Option-Impact Assessment (OIA): Before finalizing any sanction or remedy, ask systematically: *Whose choice sets will shrink or expand, by how much, and for how long?* For example, if expelling a student solves one problem but destroys that student's future options, is there a better way? Adjust until the net effect of the action *expands* the menu for those harmed and at least *preserves* as much as possible for bystanders. Ideally, even the offender's long-term menu isn't needlessly obliterated (beyond necessary restrictions for safety). This ensures we're not just doling out pain; we're managing options.
- Reversibility Index (RI): Give proposed actions a score for how easily they can be undone or dialed back. For instance, a license suspension with a re-testing pathway has a higher reversibility (better) than an irrevocable permanent ban. A moderated social media posting status that can be lifted after training is more reversible than a lifetime ban from the platform. Prefer responses high on this reversibility index.

- Repair Milestone Matrix: Lay out a matrix of observable steps for repair and track them. For example, 100% restitution completed, X hours of community service done, successful completion of a training program, apology delivered and acknowledged by the harmed party. These milestones make it clear what “completion of repair” looks like and allow all sides to see progress.
- Shadow-Price Check: Recall that in LoF, *shadow price* is like the urgency or “cost” that increases as options diminish. Apply this thinking: if a sanction’s felt burden skyrockets for someone with a shrinking horizon (say, a hefty fine on a person who’s already poor and ill is far more devastating than on someone well-off), adjust. Either add supports for that person (payment plan, fine reduction coupled with assistance) or choose a different, lower-load consequence that still protects others. The idea is not to let our justice action inadvertently impose an impossible debt on someone who has no capacity to bear it.

19.3.5 Examples (brief and concrete)

Let's illustrate restorative approaches in different contexts:

- Workplace harassment (substantiated case):
 - *Harmed employee*: Immediately ensure their safety and well-being – perhaps adjust schedules so they don’t have to interact with the harasser, provide counseling coverage, offer a role reassignment if they desire it, and compensate any lost time or opportunities (e.g. credit back sick days used due to distress).
 - *Responsible employee*: Require them to undergo mandated training (e.g. on harassment and respect), deliver a sincere apology through an approved channel (possibly in writing with HR oversight), work under monitoring on probationary terms, and set a staged plan where certain privileges (like supervisory duties) are restored only after demonstrated compliance and positive peer feedback over time.
 - *Community*: Reaffirm the company’s policies to all staff, provide coaching to supervisors on enforcing norms, open an anonymous feedback channel for others to report issues, and do quarterly “climate checks” to gauge if trust is being rebuilt in the team.
- Teen vandalism at school:
 - *Harmed parties*: The vandalized facility (say a classroom or gym) is promptly repaired using funds from restitution or supervised student

service. The club or class that lost their space gets priority scheduling for another space or extra resources as compensation for the disruption.

- *Responsible student:* They participate in a restorative circle with those affected to hear impact statements. They agree to a repayment or community service plan (like helping in maintenance or school beautification) under a mentor's guidance. Instead of a flat-out suspension that just keeps them out of school (shrinking their future), they might attend an in-school accountability program – they continue learning but with added responsibilities and check-ins to make amends.
 - *Community:* The school holds an assembly or workshop on respect for school property to reinforce norms. They might form a “student repair corps” or similar group where students help fix things and learn skills, turning bad incidents into teachable moments and involvement opportunities.
- Online misinformation (repeat offender in a community):
 - *Harmed/public:* Immediately, all instances of the misinformation are labeled with visible correction banners; links to verified information or fact-checks are boosted so that readers see accurate context. The goal is to restore truth in the information ecosystem for everyone.
 - *Responsible user:* Their content is removed, and their account is put into an education mode: for example, their posting ability is throttled and they are required to complete an accuracy or media literacy module. They might need to earn back trust by, say, contributing a certain number of verified posts or useful community notes. Only after they've demonstrated better behavior do they return to normal posting rights.
 - *Community:* The platform publishes a transparency report about the incident (without doxxing the user) and how it was handled, to maintain trust. An appeal channel is open if the user thinks they were unfairly treated. Additionally, data about the misinformation spread (de-identified) might be shared with researchers to help improve detection in the future – turning the incident into a learning opportunity for the wider fight against misinformation.

19.3.6. Metrics that matter (and can be audited)

To know if restorative justice is working, measure things like:

- Harmed-party option delta: Measure the harmed person's range of options *before* vs. *after* the intervention. For example, can the victim now go about their life (work, education, social activities) as they could before the incident, or even with fewer impediments? An increase here indicates successful restoration.
- Repair completion rate and time-to-closure: Track what percentage of the “ledger items” (losses) have been addressed and how long it takes to close each. This shows whether the process is efficiently delivering results or dragging on.
- Recurrence (with horizon control): Look at re-offense rates but adjust for baseline constraints. In other words, did similar incidents happen again, and if so, were they in contexts where maybe the original solution didn’t widen options as thought? By controlling for factors like poverty or prior history, you see if the response truly reduced *future harm* in a fair way.
- Community trust index: Survey or gauge the community’s willingness to report problems (a sign they trust the system won’t backfire on them) and their perception of the fairness and clarity of the process. If people are more willing to come forward and generally feel the system is just, that’s a good sign.
- Disparity audits: Break out outcomes by age, gender, race/ethnicity, income, disability, etc., to see if certain groups are faring worse. If, for instance, minority participants consistently get harsher outcomes or slower restorations, that’s a red flag. Adjust policies if you find disparities, so the system remains fair across the board.

19.3.7 What restoration is *not*

To avoid misunderstanding, clarify what restorative justice does not mean:

- It’s not excusing harm. Accountability is absolutely required. The person who caused harm isn’t let off the hook; rather, they are put *on* the hook in a productive way.
- It’s not “pain for pain’s sake.” We do not endorse retaliation or suffering as a *goal*. Punishment isn’t about making someone hurt because others hurt – it’s only justified insofar as it leads to repair or safety. Pure retribution (“an eye for an eye”) that doesn’t help anyone recover is off the table.
- It’s not endless process. Restoration must have an end-point: once the milestones are met and the repair is done, the stigma is lifted and the case is closed. We don’t hold someone’s past over them forever if they’ve made things

right. Likewise, the harmed party should feel a sense of closure – justice shouldn’t demand they relive it perpetually.

- It’s not metaphysical bookkeeping. We are not claiming to “balance the cosmic ledger” or that we can make everything as if the harm never happened. We focus on *concrete losses* we can repair and *real options* we can restore in the here and now, not on some spiritual notion of karma.

19.3.8 Templates you can adapt

Here are example wordings to use in agreements or closure statements:

- Restorative Agreement Header: *“This plan aims to reduce present harm, restore lost options, and prevent recurrence. It specifies actions, timelines, supports, and review dates. Completion of these terms will restore the person’s standing; non-completion will trigger proportionate alternatives.”* (This makes clear the purpose: we’re here to fix and reintegrate, not to doom anyone, but there’s accountability in failing to follow through.)
- Closure Statement: *“Milestones met. Access and standing are restored. The record reflects completion of repair. Future misconduct will be judged on its own facts.”* (This emphasizes that once it’s done, it’s done – the person is not on a perpetual blacklist, and any new incident will not be prejudged by this one.)

19.3.9 Where we go next:

Having outlined what human justice should (and shouldn’t) do under LoF, in 19.4 we turn to practical science. Next we set strict guidelines for testing LoF in simulations and AI models, ensuring we never create or tolerate suffering in silico. We’ll delve into how researchers can explore LoF’s ideas *only* in non-sentient systems.

19.4 Research Notes: Non-Sentience in Simulation

Purpose: When we build simulations to test LoF ideas – whether they are agent-based models, reinforcement learning environments, or “toy society” simulations – we must guarantee that nothing we create can *feel*. We have a moral duty to ensure our code does not accidentally produce a sentient (and thus suffering-capable) entity. This section specifies design constraints, operational tests, audits, and fail-safes so that our simulations remain strictly non-sentient while still being scientifically useful.

19.4.1 What “non-sentience” requires (operational definition)

We treat a simulated entity as non-sentient only if, throughout the entire run of the simulation, it meets all of the following conditions:

- No phenomenology-enabling architecture. The system lacks any mechanism *plausibly sufficient* for unified, self-accessing, recurrent representation – basically, nothing like a brain’s global workspace. In practice, this means no architecture that broadcasts information across modalities or maintains a persistent self-model. (e.g., we include no central observer module, no recurrent neural network that integrates experiences into a single “stream,” and no attention mechanism that binds everything together into a unified state.)
- No hedonic channel. There is *no* internal variable or process that functions as pleasure or pain for the agent. In other words, we might use reward signals for learning, but we never interpret or design them as the agent “feels.” They remain scalar optimization signals only, without being treated as a persisting internal “valence” state. If the agent can’t have an actual good or bad experience, it can’t suffer.
- No durable memory of first-person states. The simulation does not store episodic memories tagged as “what it was like for me.” Any memory in the system is either absent or is just task-related stats. There is no autobiographical memory that an entity could reflect on or experience. (If an agent can’t remember an experience as *its* experience, that’s another barrier to sentience.)
- No cross-modal integration loop. The perception→valuation→action pipeline in the simulation is strictly feed-forward, narrow, and bounded to specific tasks. There are no long-range feedback loops that could allow an agent to compare modalities or reflect on its own state (for instance, a vision module feeding into a thought module and then back into perception in a loop). These exclusions are intended to avoid constructing anything like a unified “stream of consciousness.”

- No survival/identity incentives. We do not give agents any goal related to their own continued existence or identity. For example, we never program an objective like “do not let yourself be turned off” or “preserve your memories.” An agent should never be in a position to *care* whether it’s running or not as a self. If we don’t imbue it with a self-preservation drive, we avoid a major hallmark of sentient beings.

Given the difficulty of being 100% sure about what could spark a glimmer of consciousness, our working stance is a conservative exclusion: *if in doubt, assume a design could be sentient and downgrade it* to a simpler version that is clearly non-sentient. It’s better for our simulation to be too dumb to feel than even slightly possibly feeling.

19.4.2 Design constraints for non-sentient simulators

Following from the above, here are concrete design rules for building AI agents or simulated actors in our LoF experiments:

- Stateless or very shallow agents. Favor agents that have no long-term state. For instance, use pure reactive policies: lookup tables or feed-forward neural nets *without* recurrence or long memory. If you must have learning, use episodic reinforcement learning where any internal hidden state is reset at each episode and carries over no hidden state between steps
- No global broadcast architectures. Prohibit any design that approximates a *Global Workspace*. That means no Transformer models that attend over many modalities at once, no architecture that has a central memory accessible by various subsystems, and no “controller” loop that reads and writes a shared data pool. Agents should operate in silos or very narrow pipelines.
- Myopic objectives only. Only give agents short-horizon tasks. For example, optimize immediate next-step prediction or one-step rewards, with no cumulative reward across a long episode. Do not allow discount factors that effectively create long-term utility calculations or “life goals.” We want them nearsighted in every sense.
- No affect proxies. Never label anything in the code as “pleasure” or “pain” or treat reward signals as if they were feelings. Rewards are just numbers for optimization. We should avoid even anthropomorphic language in code or comments. Don’t, for instance, have a variable called `satisfaction_level` that might tempt interpretation. Keep it strictly technical (e.g., `reward_signal`).

- Truncated memory. If the agent must have a memory buffer (for example, a frame stack in a video game), cap it at a very small number of steps ($N \leq 8$) and store only minimal statistics (averages, counts) rather than rich sequences. Absolutely no replay of past experiences allowed. We don't want an agent reflecting on or reliving experiences.
- Limited simulation scope. Model agents as extremely simple entities. A good practice is to use finite-state automata or basic Markov Decision Process policies with hand-coded transitions for multi-agent scenarios. Do not include any state variables that represent "self" or an internal narrative (e.g., avoid structures like `agent_mood` or `agent_belief_about_self`). The agent should be a collection of rules, nothing more.
- Narrative filters. If the simulation produces any output about the agent, keep it non-anthropomorphic. Do not generate text like "Agent A is sad" or "Agent B is afraid." If you need to log something, use objective terms (or just metrics). Visualize performance as graphs of reward or state counts, not as emoji faces or human-like avatars that might fool us into empathy or suggest the agent had feelings.

19.4.3 Red-team checks ("sentience dry-run")

Before running any simulation, perform a rigorous "red team" audit to catch any sentience risks:

- Architecture linting. Run an automated scan of the code/model architecture for forbidden patterns: recurrence, attention heads that span modalities, memory modules that aren't cleared, self-referential loops, etc. The lint tool should flag anything that violates our design constraints (like if someone accidentally included an LSTM or something).
- Objective audit. Search through the reward functions and auxiliary objectives in the code. Look for any long-horizon terms (like accumulated reward), any penalties or rewards that explicitly reference survival (e.g., +1 for staying alive each turn – ban that), or anything that increments based on an internal state. Ensure no "self-preservation" or identity-relevant objective is in there.
- Data-flow tracing. Analyze the computation graph to verify there's no feedback loop that would allow an output at time t to feed into a central workspace and then influence input at time $t+1$ in an unbounded way. We want to be certain that information flow is contained and can't produce a self-cognizant loop.

- Human review committee. Have a blinded committee (people who were not the designers) go over a summary of the model architecture and parameters. They then sign a Non-Sentience Declaration (NSD), which lists the model class, memory limits, horizon length, and explicitly states what the model *doesn't* have (e.g., “no global workspace, no recurrence, no self-model”). The idea is to get independent confirmation that, on paper, this thing can't think or feel in any human-like way.
- Kill-switch rehearsal. Actually test that you can terminate the simulation at any step externally. There should be a mechanism (an interrupt or a monitored condition) to hard-stop the simulation if a certain flag is raised. Demonstrate it works. Also verify through logs that the agent had no goal or method to resist shutdown (of course, we wouldn't have given it one per design, but we double-check that nothing emerges like “agent tries to save state if about to terminate”).

19.4.4 Runtime monitors and tripwires

Even with all the above, maintain live monitoring during simulations to catch any anomaly:

- Recurrent-energy monitor. Have a counter that tracks any usage of hidden state or recurrence. It should remain zero (or within some pre-approved trivial bound). If, say, someone sneaked in a recurrent connection and the energy starts accumulating, the monitor triggers an alert or shut-down.
- Horizon guard. Monitor any attempt by the program to allocate memory beyond the set buffer window or to start storing history beyond the allowed limit. If the agent tries to create a bigger memory or call a replay function, immediately interrupt and halt the run.
- Semantic filter. If any part of the system tries to label an internal variable with words like “pain,” “happy,” “fear,” etc. (for instance, maybe someone adds a logging statement “agent felt pain at time 5”), pause the run and flag a violation. This might seem extreme, but it's both practical (catching any anthropomorphic slip) and symbolic (we enforce the mindset that these terms mean stop).
- Self-reference detector. Perform continuous static or dynamic checks for any computations of the form “agent's own state influencing agent's next state because it's the *agent*.“ In code, if there's any attempt to evaluate something like `f(agentInternal_state)` where it's feeding into itself labeled as `self`, reject it. Essentially, ensure no part of the code is doing “I think, therefore I am” kind of loops.

19.4.5 Documentation and transparency (for IRBs, journals, and the public)

For each simulation study, prepare the following and make them available:

- Non-Sentience Declaration (NSD): A one-page report that anyone (ethics board, peer reviewer, or curious public) can read. It includes a diagram of the model, key parameters, memory caps, types of training data, and explicit negative statements like “No global workspace present. No recurrent networks. No persistent memory beyond 5 steps. No reward functions tied to agent survival,” etc. It’s basically the proof that we designed for non-sentience.
- Bill of Materials (BoM): Provide the exact code version, library versions, random seeds, and config files that were used. This way, anyone can reproduce exactly what we ran and verify the non-sentience claims. If we say “no recurrence,” our code is available for others to confirm that.
- Public non-sentience test harness: Release a minimal version of our guardrail tests or monitoring tools so others can run a “sentience check” on similar models. This artifact lets an external party verify that a given model class or codebase would be caught by our system if it had suspect features. It’s about building trust – not only do we say we did it right, we give others the means to double-check us.

19.4.6 Allowed vs. prohibited classes (quick reference)

Allowed (examples): These are okay types of models given our rules:

- A tabular Q-learning agent that always takes the greedy action per step, with no experience replay and a horizon of ≤5 steps. (No long-term memory, just immediate updates – very safe.)
- A convolutional neural network (CNN) policy for a gridworld game that takes the last 4 frames as input (frame-stack of 4) and no recurrence. It doesn’t carry anything over episodes and has no cross-episode memory.
- A rule-based automaton simulating multiple agents in a queue (like a checkout line) – basically a fixed state machine for each agent with no variables representing self-awareness.

Prohibited (examples): These designs are not allowed because they break one or more rules –

- A Transformer-based agent that takes in text, audio, and vision and uses cross-modal attention with an internal world-model. (This is essentially building an architecture that could integrate a lot of information like a brain might – too risky.)

- Any agent with a recurrent “controller + memory” (like an LSTM or GRU) in an open-ended task environment. If it can loop and remember arbitrarily, it’s out.
- Any model fine-tuned to report on its internal state in natural language or specifically trained to avoid being shut down. For example, an AI tuned to say “I am not a robot, I am alive” or resist termination signals is blatantly disallowed. (We absolutely forbid training like that.)

19.4.7 Why simulations still help – even without feeling

One might ask, if we strip agents of any capacity to feel, can they still tell us anything useful about LoF? Yes, they can stress-test structural aspects of the hypothesis. Non-sentient simulations can demonstrate whether certain *patterns* appear purely due to structural constraints, absent any actual experience.

For instance, we can check if algorithms that have *no feelings* nonetheless exhibit or fail to exhibit LoF-like regularities when we impose analogous guardrails. If a minimal non-sentient system *doesn’t* show the predicted patterns (like menu-narrowing near a terminal state, etc.), that strengthens the claim that those patterns in real life might specifically depend on conscious experience (i.e., QS and feelings). Conversely, if even a dumb robot shows something like “balance” under certain conditions, maybe that part of LoF isn’t uniquely tied to sentience. In short, these sims test the plausibility of LoF’s mechanisms in a controlled way.

The key is: we use simulations to probe constraints, not to demonstrate actual hedonic experience. If a constraint holds only when real feelings are in play, a non-sentient sim might *fail* to show it – which is valuable data, because it points to where consciousness itself might be playing a critical role. So simulations without feeling are not pointless; they’re our null test to ensure any alleged LoF effect isn’t just a trivial artifact of structure.

19.4.8 Reporting template (excerpt)

When we publish results from these simulations, we include a structured summary. For example:

- Claim tested: e.g. “Does horizon tightening increase selection of repair-oriented policies?” (We clearly state the hypothesis in plain language.)
- Model class: e.g. “Finite-state agents with no recurrence or shared workspace.” (So readers know up front the agent architecture was non-sentient.)

- Non-sentience audit: e.g. “NSD: Passed (architecture lint clean, objective audit clean, monitors showed no anomalies).” We reference our Non-Sentience Declaration ID and note it passed all checks.
- Results: e.g. “Observed structural pattern X with guardrails on, pattern Y without guardrails; no affective variables were present or needed.” Essentially we say what happened in terms of structure, and we remind that no feelings were involved.
- Limitations: e.g. “Simulation cannot speak to phenomenology; these results address constraint plausibility, not hedonic *reality*.” We are explicit that while the simulation supports or refutes some structural part of LoF, it doesn’t prove anything about actual conscious experience. It’s about whether the math could work, not whether life *feels* that way.

19.4.9 Failure modes (and what we do about them)

Even with precautions, things can go wrong. We catalog possible failure patterns and how to handle them:

- Silent creep toward capacity: Maybe a library update or a subtle bug introduces a bit of recurrence or memory that wasn’t intended (the agent starts to “remember” more than it should). Our monitors catch the anomaly (e.g. hidden state usage rising) and halt the run. The response: we roll back the update or fix the bug, re-run the NSD, and only then resume. Essentially, if something sneaks in that breaks rules, we *stop, revert, and re-attest* before proceeding.
- Anthropomorphic leakage: Suppose our visualization team, meaning well, labels a graph “Agent Stress Level” or uses a sad face icon for a metric. This is a communication failure – it could lead observers to think the agent had emotions. Solution: Immediately rename such outputs to neutral terms (e.g. “loss1” and “loss2” instead of “stress” or “joy”), and update our communication standard operating procedures. Everyone on the team gets a reminder: no humanizing language in outputs.
- Objective drift: Perhaps during model tuning, someone adds a new reward term that accidentally spans across episodes (a long-horizon term). If our objective audit or later review catches it (“hey, why is this reward accumulating?”), we remove it and document why anything that accumulates beyond the allowed window is disallowed. This becomes an example in our internal knowledge base to prevent similar mistakes.

19.4.10 Ethics summary for simulation work

To wrap up our stance:

- We only simulate structures and behaviors, never experiences. (If it looks like we might simulate an experience, we stop.)
- We never equate an AI's reward signals or internal variables with pleasure or pain. Those are just numbers, not feelings, and we treat them as such.
- We publish our non-sentience proof (the NSD and related materials) with every simulation study so that others can verify our ethical compliance.
- If future science tightens the definition of what kinds of architectures *could* support sentience, we will tighten our exclusions accordingly. (Our policy ratchets in one direction: more caution as knowledge advances.)

19.4.11 Where we go next:

In 19.5 we leave the lab and turn to the real world of communication. Next we discuss how to talk about LoF responsibly — in classrooms, clinics, the media, and online — so that our words remain accurate, compassionate, and safe for all who hear them.

19.5 Communication Ethics

Purpose: The Law of Fairness intersects with deeply human matters – pain, hope, grief, meaning-making. This means that how we communicate about LoF must be handled with great care. Our communication needs to be accurate, appropriately cautious, compassionate, and safe for our audience. In this section, we set standards for how to speak and write about LoF in various arenas: classrooms, clinics, labs, the media, and online communities.

19.5.1 First principles

These guiding rules apply to all communication about LoF:

- Do no harm in words. Avoid any language that could worsen someone's distress, stigmatize those suffering, or encourage harmful behavior. Words can wound. We choose them as if a vulnerable person is in earshot – because they likely are.
- Truth over allure. It's tempting to make grand, sweeping statements to capture attention, but we must favor precise claims with clear uncertainty markers over dramatic proclamations. If a result is preliminary, say so plainly – don't dress it up as definitive for the sake of a headline or inspiring quote.
- Respect persons. People are not data points or test subjects in conversation. When telling stories or presenting cases, anonymize appropriately and highlight dignity and agency. No slide or social media post should ever treat an individual's life lightly. Always ask: am I treating this person (or group) with the same respect I'd want for myself or my loved ones?
- Separate hypotheses from guarantees. LoF is *not* a promise that "everything happens for a reason" or that "your suffering will be rewarded." It's a scientific hypothesis about constraints on experience. Always delineate what is an unproven idea (however interesting) versus what is solid evidence. Do not let an audience confuse LoF with some moral guarantee of goodness or fairness – it's not that.

19.5.2 Talking about pain, suffering, and suicide – safely

These are the heaviest topics connected to LoF, and there are well-established guidelines (which we adopt) to discuss them safely:

- Person-first, nonjudgmental language. Always talk about individuals as people *with* experiences, not as labels. Say "person *living with* depression" rather than "a depressed person," which can sound defining or reducing. When discussing

suicide, say “died by suicide” instead of “committed suicide” (which carries criminal connotations or blame). This keeps the focus on compassion rather than judgment.

- No glamorization or simplification of suicide. Never frame suicide as a logical or even quasi-positive outcome, and absolutely do not imply LoF provides an “explanation” or excuse for it. For example, do not write a story that implies “In LoF’s view, maybe it made sense this person died.” That is dangerous and false. LoF never justifies self-harm; any suggestion otherwise is strictly forbidden.
- Include help-seeking pathways. Whenever content might be triggering (discussing severe suffering, despair, etc.), *always* close or accompany it with information on how to get help. This could be a general statement like “If you are struggling, please reach out to a mental health professional or crisis line – help is available,” plus a link or number for a local resource. We want to gently guide anyone in the audience who feels personally affected toward safety.
- Model uncertainty *and* care. If someone in the audience is suffering, our message should be: “We *don’t* know your individual situation fully, but we do know there are people who care and want to help; you matter.” It’s important to avoid any deterministic tone like “LoF says your pain is fine.” Instead, emphasize support and that research is ongoing – we’re seeking answers, we don’t have them all, but regardless, the person’s life is valuable.
- Content advisories. If a talk or section of content will delve into trauma, grief, or end-of-life issues, warn your audience beforehand in a respectful way. E.g., “Note: The next section discusses end-of-life experiences and suicide. It may be difficult for some; please care for yourself and feel free to step out if needed.” This respects individual boundaries and triggers. It’s not about slapping a sensational “viewer discretion”; it’s a gentle heads-up so people can make informed decisions for their own well-being.
- Example safe script for a public talk: *“Today we will discuss some findings about end-of-life experiences and how scientists measure feeling. If this topic is hard for you, please take care of yourself – you can step away at any time. And remember, if you’re struggling or having thoughts of suicide, you’re not alone and there are people who want to help – consider reaching out to a trusted professional or local support service.”* This way, we prepare the audience and weave in a lifeline in case someone is on the edge.

19.5.3 Avoiding overclaiming (and how to correct if it happens)

- Label the scope of every claim. We will tag our statements clearly: Is it a *hypothesis*? A *pre-registered prediction*? A *replicated finding*? Or just a *speculative analogy*? For instance, instead of saying “LoF shows that dying people feel X,” say “Our *hypothesis* (not yet proven) is that as horizons shrink, people might feel X.” If we have data: “In one small study (preliminary), people near end-of-life *reported* X.” This way, listeners always know how much weight to give a statement.
- State limitations explicitly. If our data is correlational, say exactly that: “This is correlation, not proof of causation.” If N is small: “Sample size was small, so we’re cautious.” If all our participants were e.g. from one country: “We don’t know if this generalizes culturally.” Being upfront about limitations builds credibility and prevents misinterpretation. We do *not* hide caveats in footnotes – we put them right in the narrative.
- Have a fast correction policy. If despite our best efforts we or our affiliates misstate something or new evidence overturns what we said, we will correct it *publicly and promptly*. For example, if a press release claimed “LoF effect confirmed” and later the replication fails, we issue a correction on our website, social channels, etc., within 72 hours, clearly stating the new understanding. In papers or preprints, if an error is discovered, we publish a formal erratum or update. The idea is to own mistakes and model intellectual honesty.

19.5.4 Evidence-forward storytelling (without anthropomorphism)

When explaining LoF concepts, we often use stories or analogies. That’s fine, but we must tether them to evidence and avoid sneaky teleology:

- Show structure, not inner life. When describing QS or horizon effects, frame them as changes in available options or measurable patterns, *not* as a little agent in your head “wanting fairness.” For instance, say “As a crisis looms (horizon shrinks), people’s choices naturally narrow toward immediate relief (the data shows more short-term decisions)” rather than “Your brain wants to balance the scales.” We avoid imputing desires or consciousness to the system – we describe what it *does* in structural terms.
- Use visuals that don’t mislead. Prefer diagrams of processes (flowcharts of QS mechanism, graphs of affect over time with confidence bands, etc.) and real data plots over, say, brain scans with hot spots (which can be overly dramatic or imply we found “the fairness center in the brain” nonsense). If we show a brain image, it

should be properly annotated and dull-colored, not the neon brain porn people misuse. Code snippets or equations (with explanation) can sometimes convey rigor better than some glossy stock image. The key is not to inadvertently hype with visuals – they should illuminate, not inflate.

- Anecdotes ≠ data. We might start a talk with a patient’s story or a hypothetical scenario to motivate a question, but we must always loop back to “So we designed a study to test this,” or “This story is an illustration; here’s what our evidence says.” We never present an anecdote as proof. Stories are doorways to understanding a problem, not answers. For every narrative we give, we pair it with the actual measurement plan or falsifier it inspired. That way, listeners see that we move from anecdote to analysis, not the other way around.

19.5.5 Public conversations and moderation (classrooms, forums, social media)

When hosting or participating in discussions about LoF in public or educational forums:

- Set clear house rules. If it’s our forum (like a subreddit we moderate or a Q&A after a lecture), we explicitly state the boundaries: no one should solicit or dispense personal medical advice there (“we are not doctors here, seek professional care”), absolutely no encouraging self-harm or fatalism, no doxxing or personal attacks, etc. We frame it as “We’re here to discuss ideas, not to endanger or belittle anyone.”
- Compassionate redirection for personal crises. Inevitably, someone might share they are in a dark place or something very personal. Our duty is to acknowledge their experience, thank them for sharing/trusting, and gently redirect them to appropriate help, *not* to engage in a theoretical debate about their condition. If possible, take it off the public thread (“I’m going to DM you some resources” or “Perhaps we can talk after class with the counselor”). The aim is to protect their privacy and get them help, not leave them vulnerable in a public arena.
- Respond to misinformation with facts, not scorn. If someone posts a misconception (“LoF proves karma is real!” or “This is just like The Secret”), we correct it factually and maybe provide a reference: “Actually, no – LoF is just a hypothesis about summed experience and has *no* evidence for cosmic justice.” We do it without belittling the person. Sarcasm or elitism will just drive people away or entrench false beliefs.
- Cultural humility. Encourage and welcome perspectives from people of different backgrounds or faiths. If someone says “In my culture, suffering is seen as...,” we listen and acknowledge that LoF is a scientific lens, not a replacement for

personal or cultural values. Make it clear we're not here to declare centuries of philosophy or religion obsolete. LoF might intersect with those views, but it isn't *above* them. We position it as one way of looking at things, grounded in data, and open to critique from ethical and cultural vantage points.

19.5.6 Media and policy engagement

When interacting with journalists or policymakers, we follow special precautions:

- Maintain a one-page fact sheet (and share it). We keep an updated document that any reporter or official can have which states: the canonical LoF statement, *what LoF does not say* (e.g. “does not say life is fair or that you should endure pain”), the current level of evidence (e.g. “2 correlational studies, no causal proof yet”), and open questions under investigation. This preempts a lot of misunderstanding. Instead of them Googling random pieces, we give them an authoritative summary.
- Use disciplined sound-bites. We craft and rehearse a couple of short, *correct* descriptions for interviews. For example: “*The Law of Fairness is a hypothesis about a constraint – it predicts patterns in how experiences add up, and we’re testing those patterns.*” And we explicitly avoid sexy-but-wrong phrases like “Life is truly fair, science shows!” If a journalist tries to get a punchy quote like “So, this proves karma?”, our answer is something like: “No, it doesn’t prove anything metaphysical – it’s a research framework that might explain certain psychological patterns. It’s testable and could well be false in part.” We give them a clear takeaway without hype.
- Policy guardrails. If we are consulted on or advocating for any policy, we *never* support one that inflicts short-term harm on some group because “LoF says it’ll even out later.” For example, if someone suggests, “Maybe we can justify cutting mental health services because people will compensate,” we vehemently shut that down: LoF is *not* utilitarian arithmetic, and it cannot be used to justify reducing anyone’s well-being now for a speculative later payoff. Our rule in policy is: LoF insights can inform ways to broaden options and improve support, not to justify neglect or exploitation. We will always push for policies that protect the vulnerable and preserve future options, in line with our guardrails.

19.5.7 Ethics of simulation and AI communication

A special subset of communication ethics deals with how we talk about our simulations and AIs (given the pitfalls we discussed in 19.4):

- Non-sentience commitments front and center. Every result from a simulation should be accompanied (in the paper or press release) by a statement like: “All agents in this simulation were certified non-sentient (see the Non-Sentience Declaration (NSD) for this run).” We proactively disclose the steps we took. Not only is this transparent, but it also educates the audience about these precautions as a norm.
- No anthropomorphic dashboards in public. When we present our AI’s performance, we use neutral language. For instance, instead of “The AI got frustrated when...”, we’d say “The AI’s error metric increased when...”. We avoid showing any interface or graphic that suggests the AI has emotions or a persona (unless we’re explicitly making a point about such anthropomorphism as a problem!). Essentially, we train ourselves and our visuals to refrain from labeling AI internals with human feelings.
- Open-source materials, careful framing. Whenever possible, we release the code or configuration that shows how we kept the AI non-sentient. This invites trust and external verification (it also helps spread best practices). At the same time, when framing the results, we remind everyone: this simulation’s results are about structural plausibility, not evidence of conscious experience. E.g., “The code and config are available online so you can verify these agents had no capacity for feeling. The patterns we observed suggest how certain constraints might operate in principle, but they do not demonstrate any actual subjective experience.”

19.5.8 Sample scripts you can reuse

Sometimes it helps to have a ready phrasing for tricky interactions:

- A. Gratitude + boundaries (replying to a thoughtful but vulnerable community post): *“Thank you for engaging so thoughtfully with the Law of Fairness and for keeping the conversation respectful. Just a gentle reminder: LoF never justifies neglecting real pain, and it never endorses self-harm. If this topic is bringing up difficult feelings, please step back and consider talking with a qualified professional or a local support service. We’ll keep our discussion focused on the science, with care for everyone here.”* (This message does several things: it thanks them, sets a boundary that we won’t go into harmful territory, offers a path to help, and refocuses on science.)
- B. Press inquiry (short form answer to “What is LoF?”): *“The Law of Fairness is a hypothesis about a constraint on conscious experience – it’s not a promise or a cosmic law. We’re basically testing whether certain measurable patterns hold,*

like how a person's options change as their time runs short. And we're very clear about the limits: our data is early and this could be disproven. It's an idea we're checking with experiments, not a final truth." (This gives them a quotable explanation that emphasizes testability and caution.)

- C. Public correction (for when we ourselves overstep in a post or talk):
"Yesterday I overstated the evidence for horizon effects in end-of-life experiences. To correct that: we have some early signals, but no replicated, cause-proven result yet. I've updated the online post and linked to our preregistered study that's in progress. Thank you to those who pointed out the overstatement – holding us to a high standard helps everyone." (This exemplifies humility and responsiveness, and it directs folks to the actual evidence.)

19.5.9 Closing Ethic

We aim to speak carefully, measure honestly, and care publicly. Before publishing, presenting, or posting anything, we ask ourselves: "Does this help a thoughtful reader become clearer, kinder, and safer?" If not, we revise it – or we don't publish it at all. No result, no paper, no tweet is worth contributing to confusion, cruelty, or danger. Clarity, compassion, and safety are the north stars by which we navigate all discourse around LoF.

19.5.10 Where we go next:

By drawing these rigorous boundaries around research, we've prepared the ground for one final ethical checkpoint. In 19.6, we compile the absolute *hard lines* that no one in this project will ever cross. These are our inviolable red lines — a summary of non-negotiables to ensure that no pursuit of LoF ever violates core human dignity.

19.6 Hard Lines We Will Not Cross

Purpose: LoF is a scientific hypothesis about constraints on conscious experience – not a free pass to do whatever we want in its name. This section enumerates the unambiguous red lines that our team (and any ethical practitioner of this research) will not cross in research, communication, product design, or community practice. These are non-negotiable. If any of these lines is at risk of being crossed, the work stops. Period.

19.6.1 Non-negotiables (one-glance list)

- No “harm justification.” We will never justify causing someone pain or harm now by appealing to a hypothetical long-run neutrality or balance. (*In practice: you cannot say “It’s okay to hurt X because eventually it evens out” under LoF.*)
- No self-harm encouragement. We will never encourage, rationalize, or gamify suicide or self-injury. Not in a hint, not indirectly – never. If someone is suffering, the message is to get help, not to suggest ending their life or that their pain is somehow fine.
- No exploitation. We never target or pressure vulnerable groups to bear burdens “for the sake of data” or any supposed greater good. For example, we won’t recruit financially strapped people into risky experiments by dangling money, or push end-of-life patients to endure more for science.
- No sentient simulations. We do not knowingly build or run any AI or simulation that is *plausibly capable of suffering or consciousness*. If there’s a credible chance a system could feel, we either redesign it to eliminate that chance or we do not run it at all.
- No privacy breaches. We will not collect sensitive personal data without clear consent. We will not retain it longer than necessary. We will not attempt to re-identify anonymized data. Participant privacy and control over their info are inviolable.
- No overclaiming. We will not present hypotheses as if they are proven facts, and we won’t hide uncertainty to get attention or funding. In grant proposals, press releases, or any outreach, we’ll stick to what the evidence actually supports.
- Stop-work rule. If any credible report arises suggesting that what we’re doing is creating a risk to someone’s safety, dignity, or rights, we hit pause. Work stops, and we review. We’d rather lose time or money than cross an ethical line unaware.

19.6.2 Human research red lines

- No deception about risks. In our consent process, we will not downplay or hide risks. No burying discomfort in fine print. Participants must be fully informed of any potential physical, emotional, or privacy risks in understandable language.
- No inducing high distress for data. We will not design studies that intentionally push participants into more than minimal risk levels of suffering just to observe them, even if someone argues it would yield “valuable insight.” For example, we wouldn’t deprive someone of sleep to extreme levels or expose them to trauma triggers beyond what clinical practice allows – not acceptable.
- No leveraging care access. We will never condition someone’s needed care or benefits on participating in our research. e.g. We won’t say “You can only get this therapy if you enroll in our study” or exploit someone’s reliance on services to coerce them into an experiment. Participation is entirely voluntary and not tied to medical care, financial aid, housing, grades, employment, etc.
- No terminal manipulation. End-of-life research must be strictly observational and *comfort-first*. We would never withhold standard pain relief, spiritual support, or family visitation just to get a “clean” measurement of final days. The dying process is not a lab exercise; patient comfort and dignity trump all.
- Community veto power. For studies involving populations like hospice patients, trauma survivors, etc., we commit to giving a say to representatives of those populations (e.g. patient advocates) in setting what is off-limits. If they say a certain procedure or question is unacceptable, we abide by that veto. The people most affected get a final cut on the ethics.

19.6.3 Animal research red lines

(While LoF work so far is human-centric, if it ever extends to animal models, these apply.)

- No pain without direct clinical value. We will not subject animals to more than transient pain or distress unless it’s *directly* tied to finding a treatment or intervention that could relieve suffering in the context of LoF (and even then, we must have no alternative methods). Exploratory “let’s see what happens if we torture this animal” is obviously out. Even standard pain-inducing paradigms must clear the bars of necessity and irreplaceability by non-animal means.
- Refinement by default. If any animals are used, we automatically include analgesia (pain relief), enrichment (to reduce stress/boredom), and set early humane endpoints (stop criteria when an animal is suffering too much or not

recovering). We also design our studies to use the minimum number of animals needed (statistical power calculated to avoid any more subjects than absolutely necessary). We never try to maximize data by just adding conditions if it increases animal use or pain beyond necessity.

- Prohibited paradigms. Some kinds of animal experiments are flat-out banned for LoF research: we will not do learned helplessness studies (where animals are given inescapable shock or failure – a classic but cruel paradigm), we won’t do extreme isolation rearing of animals, nor inescapable chronic stress models. Those might yield data on “despair” etc., but they are fundamentally at odds with the spirit of our ethics. We’ll find humane alternatives or accept not knowing those things.

19.6.4 Data and privacy red lines

- Minimal capture. We will only capture data that our analysis truly requires. By default, we try to process data on the device or on the edge and only store derived metrics (e.g. count of steps instead of the raw GPS tracks of a person). If we can answer a question with lower-resolution or aggregated data, we’ll never take high-resolution personal data “just in case.”
- No secondary fishing expeditions. Data given to us for one purpose will not be quietly reused for another without getting new consent. We also will not sell data or share it with third parties for their own use, unless participants explicitly opt in (and even then, only if it’s something good for them). Basically, your data won’t wander off to unknown places for unknown reasons.
- Anonymity with teeth. All personal identifiers are stripped early (ideally at collection) and replaced with codes. We apply techniques like differential privacy where appropriate so individual contributions are masked even in aggregate results. We regularly attempt to “attack” our own datasets to ensure individuals can’t be re-identified – and if they can, we strengthen our protections. Raw identifiers (names, emails, etc.) are deleted as soon as they’re not absolutely needed, typically right at ingestion.
- Right to vanish. If a participant withdraws or after a study if they change their mind, we will delete their data from all our storage and analyses to the best of our ability (and confirm it in backups, derived sets, etc.). We won’t cling to someone’s data against their wishes. And we make this option clear to them – it’s not hidden.

19.6.5 Communication and community red lines

- No medical advice in public forums. In any forum we moderate (online community, etc.), we will never allow ourselves or others to give personal medical or mental health advice as if we're professionals treating someone. Instead, moderators are instructed to thank the person for sharing, acknowledge their pain, and redirect them to appropriate professional help. Public forums are not safe for therapy, and we enforce that boundary strictly.
- No glamorization of crisis. We avoid telling stories in ways that make them sound romantic or heroic to be suffering. We won't detail suicide methods or dramatic self-harm incidents in our communications. There's a known phenomenon of such details triggering vulnerable individuals – we refuse to contribute to that. Our case examples will always be framed with discretion and respect, not sensationalism.
- Corrections are mandatory. If we (or community members under our purview) spread a piece of misinformation or an exaggerated claim about LoF, we commit to issue a correction in the same venue within a defined timeframe (say 72 hours for something online). We moderate our own content rigorously for factual accuracy and promptly fix any mistakes. Transparency builds trust.
- Respectful dissent. We do not silence or ban people for disagreeing with LoF or criticizing our work – as long as they do so civilly. In fact, we welcome good-faith critique. We will moderate insults, hate, or trolling, but not mere disagreement. If someone says "I think LoF is flawed," that's fine. We engage or let it be. Our rule: we moderate *tone*, not *conclusions*. Critics with substantive points are invited into the dialogue (perhaps even to collaborate or provide an adversarial review), not chased away.

19.6.6 Simulation and AI red lines

- Non-Sentience Declarations (NSDs) required. Every simulation or AI experiment we release or publish must include a Non-Sentience Declaration that describes why the system *cannot* be conscious or feel. If an AI project doesn't have an NSD signed off by independent experts, we simply don't run or publish it. This is as crucial as an IRB approval for human studies.
- Kill-switches and caps in all code. All our AI systems have hardwired runtime limits (so they can't run unchecked forever) and memory/computation caps to prevent unexpected complexity growth. They "fail closed," meaning if something

goes wrong or time is up, they shut down safely by default. No open-ended self-improvement loops without oversight.

- No “pain” variables or anthropomorphic labeling. We never name an AI’s internal state things like “pain,” “anger,” or similar. Even if we simulate a reward signal for something negative, we call it, say, “negative_reward,” not “pain_score.” By policy, using such loaded terms in code or documentation is forbidden – it blurs the line and invites dangerous thinking.
- External ethics audit for AI. Before any major AI experiment, we require an external ethicist or review board to verify that the system design cannot produce valenced experience. If we can’t secure such an audit or if the auditors aren’t satisfied, the experiment doesn’t run. We treat this as seriously as biosafety for a pathogen study. No independent green light, no go.

19.6.7 Governance and conflicts

- Independence in oversight. Any board or committee overseeing our ethics (for human or AI research) will include members who have no financial stake in whether our results are positive or sensational. For instance, we might include a hospice nurse, a patient advocate, or an academic ethicist with no ties to our grants. This ensures decisions are driven by ethics, not profit or career incentives.
- Conflict of interest disclosures. All investigators, moderators, and relevant vendors must fully disclose any funding sources, equity, or relationships that could influence them. If we’re funded by, say, a tech company that might benefit from a certain finding, we lay that out openly in publications and talks. Sunlight is the best disinfectant for trust.
- Whistleblower protection. We establish confidential channels for anyone (team members, participants, outside observers) to report ethical concerns or rule violations. We have a strict no-retaliation policy – meaning if someone flags a problem, their standing (job, participation, etc.) is not threatened. Substantiated concerns will be made public along with our resolution steps. In short, we encourage people to speak up if they see something off, and we protect them when they do.

19.6.8 Crisis scenarios – pre-committed responses

We pre-load our reaction to certain foreseeable crises so we don’t make decisions in panic or self-interest:

- Suicidality disclosures (during study or online discussion): As soon as someone expresses suicidal thoughts, we *pause all other objectives*. The response is: acknowledge their pain and bravery in speaking up, thank them for telling someone, provide them with immediate resources (like “Here is help available right now”), and *move the conversation off the research/topic and into a support mode*. If it’s in a study context, research stops and clinical care begins. We never, ever engage in any theoretical discussion with someone in crisis (e.g., debating LoF or meaning of life). The standing order is care first, theory never in a crisis.
- Data breach: If any personal data we hold is compromised, we will not cover it up. Our plan: notify all affected participants promptly (as soon as we know what was taken, we tell them – no waiting weeks “to investigate fully” if people can act to protect themselves now). We also tell them exactly what we’re doing to remediate (e.g. offering credit monitoring, informing authorities if needed). We pause relevant data collection until we fix the vulnerability. And we commission an *external forensic review* to audit what happened and ensure it’s truly resolved. Trust is key; we show that by responding aggressively to any breach.
- Evidence reversal: If a preregistered falsification test fails – meaning we find evidence *against* LoF – or any major result is overturned, we will publish that negative finding. We won’t hide it or quietly drop that line of inquiry. For example, if Experiment X was meant to verify a key LoF prediction and it comes up null or opposite, we get that out in the open. We update our claims accordingly, even if it’s uncomfortable or undermines prior statements. We refuse to bury null results or inconvenient data.

19.6.9 What we refuse to say (and what we say instead)

To make these lines absolutely concrete, here are specific phrases or attitudes that are banned, paired with the correct stance:

- Refuse: “Suffering is fine because it all balances in the end.” Say: “Nothing in LoF makes *current* pain acceptable. Our duty is always to reduce harm now – any future balance doesn’t lessen today’s responsibility.”
- Refuse: “LoF explains why someone *should* end their life.” Say: “LoF never endorses self-harm. If you are hurting, LoF is not a reason to give up – please reach out to qualified help; *you matter right now*.”
- Refuse: “LoF proves life is fair.” Say: “LoF is a testable constraint hypothesis – we are actively trying to falsify it. It does *not* claim the universe is fair or just, only that we might detect certain patterns if it’s true.”

- Refuse: “We need more pain to get stronger data.” Say: “If a protocol would increase suffering beyond minimal risk, we will not do it. We’ll find ethical ways or accept the limits of knowledge. Data is not more important than people’s well-being.”

By setting these as call-and-response, we train ourselves and our community to catch and correct any drift toward dangerous interpretations.

19.6.10 Enforcement

how these lines hold: It’s easy to list rules; the hard part is enforcing them consistently. Here’s how we ensure these lines are real:

- Signed adherence. Every team member, collaborator, and community moderator signs off on this red-line policy (and renews that pledge annually). This isn’t a one-time glance – it’s a formal commitment that they understand and will uphold these rules. New members are onboarded with training on these exact lines.
- Pre-study ethics docket. For each project, before it begins, we compile a one-page “Ethics First” docket that outlines potential risks, how we mitigate them, any stopping rules, and which community representatives were consulted. This docket must be approved (internally, and by any external board if applicable) *before* research starts. It’s our way of forcing ourselves to think through ethical aspects in advance every single time.
- Periodic audits. Twice a year, we have an audit (could involve external folks or internal rotating roles) that checks: consent forms, data handling logs, simulation configs, moderation logs from forums, etc., against these policies. We then publish a summary of these audits publicly. Even if nothing went wrong, we say “We checked X, Y, Z – all clear” or “We found minor issue W and fixed it.” Regular auditing creates accountability beyond any one person’s vigilance.
- Sanctions for violations. If a violation does occur, what happens? We’ve set it that any such event results in immediate suspension of the study or activity in question, a public notice of what happened (we won’t sweep it under the rug), and, when relevant, reporting to oversight bodies (institutional review boards, journals if it affected a publication, funding agencies, or platform authorities). Individuals who willfully violate core ethics may be removed from projects or reported to professional boards. Essentially, crossing a red line triggers tangible consequences – not a slap on the wrist, but measures that matter.

19.6.11 The moral of the method

Ultimately, the way we conduct this work must embody the fairness we study. If exploring the Law of Fairness ever requires us to do something fundamentally unfair or inhumane, then we won't do it. A law about conscious experience that forgets the person in front of us has already failed – no matter how elegant its equations or results.

No “forbidden experiments” that inflict harm on sentient beings (e.g., lethal mazes or deprivation-loaded tests), including attrition by design, time-to-death horizons, or adversarial deprivation meant to coerce choices. We will develop only ethical, non-sentient, or observational proxies.

In summary, these hard lines are here to keep us honest: to ensure that in pursuing knowledge, we do not lose our humanity. If honoring these lines slows us down or limits what we can test, so be it. Integrity comes first. We will find other creative ways to investigate, but we will not cross these boundaries. The story of LoF – if it is to be written – will be one we can be proud of not just for its findings, but for the way we arrived at them.

19.6.12 Where we go next:

With our ethical guardrails firmly established, we transition to Chapter 20. In the next chapter, we bring the discussion back to personal life. We will explore how, even if LoF imposes constraints, people can still find hope, exercise genuine freedom of choice, and create meaning in their daily lives. Chapter 20 asks: How can we live fully and freely under (or without) the Law of Fairness?

Chapter 20 — Hope, Freedom, and Daily Life

Imagine you're talking with a close friend about this theory over coffee. "So you're telling me," they say, "that no matter what I do, all my happiness and sadness will even out by the time I die?" It's a disorienting thought. If you've been riding a wave of good fortune, you might worry a downturn is inevitable to "balance" things. If you're in the depths of misery, you might take comfort (or desperation) in the idea that brighter days *must* be ahead. The Law of Fairness provokes deep questions: Does it make our choices meaningless or actually more meaningful? Should it give us comfort during hard times, or does it risk turning life into a zero-sum game? This chapter grapples with those questions, exploring how LoF touches our sense of hope, our feeling of freedom, and the texture of our everyday lives.

Let's start with hope. On the face of it, saying "every life ends up emotionally neutral" might sound bleak – as if all our joys and triumphs are erased. But that's a misunderstanding. LoF doesn't erase the highs and lows of life; it stitches them together. In fact, it can be reassuring: no matter how dark a moment is, under LoF it cannot extend indefinitely or define the totality of your life. If you're suffering now, the theory implies there *must* be relief by the end – the bad times won't accumulate without bound. Many wisdom traditions echo this idea with sayings like "This too shall pass."

Now consider freedom. Does LoF mean our life story is pre-scripted to balance out, and thus out of our hands? Not at all. While the final tally may be constrained, the path to that balance is incredibly open-ended. Think of it this way: there are countless life trajectories that could all end in a neutral ledger, and your choices help determine which trajectory is yours. You remain the author of your life. The law doesn't dictate whether you become a musician or a doctor, or whether you love skydiving or prefer quiet walks – those decisions are yours, and they shape the sequence of experiences you have. LoF simply suggests that, however events play out, the grand total of felt positives and negatives will balance. Importantly, this doesn't diminish moral responsibility or personal agency. We still face real decisions about how to treat others and ourselves. If anything, LoF underscores responsibility: since everyone's joys and pains sum out in the end, the kindness or harm we offer someone won't "tilt" a cosmic scale permanently, but it absolutely matters to *that person in the moment*. We can choose to make someone's day better or worse, knowing that the immediate experience is genuine even if the long arc of their life has a balancing mechanism. In short, LoF doesn't make us puppets of fate; it just adds a background constraint – much like knowing the planet's gravity will eventually pull you back down if you leap, yet you're free to jump and dance as you wish.

How does this perspective play out in daily life? It can actually enrich it. First, understanding that highs and lows are two sides of the same coin can help us savor the good times without fearing them. Yes, extreme happiness might be followed by a downturn – not because the universe punishes you, but because that's how dynamic systems often work (your emotions recalibrate, external circumstances change, etc.). Knowing this, you might cherish joy when it comes, mindful that it's part of a larger ebb-and-flow. Conversely, in painful times, remembering LoF can be a source of strength. You might think of your emotional life like a pendulum: when it swings far in one direction, it's poised to swing back. This isn't just wishful thinking – it's how resilient minds and bodies often respond. Many people have observed that after a period of deep grief, even small comforts or kindnesses can ignite profound relief and gratitude. LoF frames such rebounds not as random twists of fate, but as an expected part of our psychological economy.

Seeing life as inherently balanced can also guard against both complacency and despair. It's not that nothing matters – it's that *everything* matters for the moment it's in. Imagine your life as a financial ledger where you'll ultimately break even. That doesn't render your earnings and spendings pointless; rather, it focuses you on what you value. If you know that splurging on reckless pleasures might incur a debt of pain later, you might aim for a steadier kind of happiness. And if you've been dealt heavy sorrows early on, you might find motivation in the idea that you have a lot of "growth" or positive experiences still in your account to discover. Crucially, the neutrality at the end doesn't nullify the journey – a rollercoaster ride ends where it started, but the thrills and lessons along the way are very real. In the same sense, a life that sums to zero isn't a life of nothingness; it's a life where every joy and every sorrow had its time and left its mark. The memories, growth, love, and wisdom gained from those experiences remain with you, even if the raw pleasure-pain ledger resets.

This chapter encourages embracing LoF's insight in a practical, life-affirming way. Rather than viewing it as a fatalistic verdict, you can treat it as a call to balance and awareness. It can be liberating to realize that you don't have to chase eternal bliss or fear everlasting despair – neither extreme will define you forever. What matters is how you navigate the swings. You might even turn this into a personal experiment: pay attention to your ups and downs and notice if and how they balance out over time. The goal isn't to force a balance (nature seems to handle that on its own), but to cultivate a kind of emotional flexibility and gratitude. In daily life, that means celebrating joy without clinging to it, and enduring hardship with the confidence that it won't last indefinitely. Hope, under LoF, is not naive optimism that "only good things will happen" – it's a resilient understanding that even when bad things happen, they contribute to a meaningful whole and will eventually

give way to better times. And freedom, under LoF, is the freedom to write the story of your highs and lows in a way that reflects who you truly are.

What you'll get from this Chapter:

- A nuanced outlook on hope: why believing in an eventual balance can be comforting in dark times without diminishing the value of happy moments.
- Reassurance about free will and choice: how a fixed end-of-life balance does *not* mean life is predestined or our decisions don't matter.
- Practical insights for daily living: seeing joys and sorrows as complementary, learning to ride life's ups and downs with more mindfulness and less fear.
- Analogies and evidence that put LoF in perspective (from emotional "budgets" to end-of-life calm), helping you integrate the concept into your understanding of a life well-lived.
- An invitation to personal reflection rather than blind belief – encouraging you to observe patterns in your own experience and consider testing the law's ideas in gentle, everyday ways.

Subsections in this Chapter:

- **20.1 Freedom Inside Guardrails** – Does a fixed fairness constraint undermine free will? This subsection argues that LoF's constraints limit outcomes, not agency. Even if certain extremes are off-limits, individuals still face real choices in their lives. In fact, knowing that some guardrails exist (if they do) can reduce panic and despair in tough times and help people make wiser decisions within those safe bounds.
- **20.2 Meaning Without Illusions** – If life is constrained to hedonic neutrality at the end, does anything we do matter? Here we contend that meaning does not require a cosmic purpose or unlimited outcomes. Human purpose can arise from the effort to restore balance in ourselves and others. Acts of practical care—reducing the “open tickets” of suffering—give life profound purpose, showing that life can be deeply meaningful even without grand cosmic narratives.
- **20.3 If the Law Is True** – How should we feel if LoF actually holds? We discuss the attitude of cautious optimism and resilience that could come with believing the law is real. If life naturally balances out, one might embrace palliative care and social policies that *work with* this balancing process (rather than fighting it), finding comfort that extreme injustices won't last forever. Importantly, a true LoF

wouldn't mean we do nothing; it would mean we cooperate with nature's balancing by being kind and patient, knowing every bit of relief helps.

- **20.4 If the Law Is False** – What if LoF is wrong? Paradoxically, the failure of LoF would increase our responsibility to alleviate suffering, not decrease it. If there is no automatic counterweight in life, then preventing and relieving pain falls entirely on us. This subsection frames that stance as actionable hope rather than despair: precisely because there's no guarantee of balance, we must strive to create as much relief and goodness as possible.
- **20.5 Talking with Skeptics** – Not everyone will buy into LoF, and that's healthy. This part offers guidance on engaging skeptics in honest dialogue. We emphasize presenting LoF as a *testable constraint*, not a dogma, and inviting critics to examine the data or collaborate on tests. By welcoming disconfirming evidence and maintaining scientific humility, we demonstrate that our aim is truth, not ideological conversion. The tone in these conversations should always be one of openness and mutual respect.
- **20.6 A One-Page Summary to Share** – Finally, we provide a concise handout distilling the entire LoF project. This one-page summary lays out the core claim, key evidence plans, and ethical commitments in plain language. It's designed as a tool readers can use to share the idea responsibly with friends, family, or colleagues. In one sheet, someone will be able to understand what LoF proposes, how we're testing it, and the care we're taking to do it ethically.

Where we go next:

We now move into 20.1, examining how personal freedom operates even if life's outcomes are bounded by LoF. In the next section, we'll see why acknowledging natural guardrails doesn't diminish our ability to choose — in fact, it can empower us to act more wisely within those limits. The journey into hope and freedom begins with understanding the space we have *inside* the guardrails.

20.1 Freedom Inside Guardrails

Thesis: The Law of Fairness doesn't shrink freedom; it *stabilizes* it. What might look like a guardrail from the outside is, from the inside, the very condition that keeps real choice available *tomorrow*. Freedom that burns the bridge behind you isn't freedom at all — it's a short-lived permission slip. Freedom that you can keep is the whole point.

20.1.1 Two kinds of freedom

1. “Brittle” freedom – *anything goes, for now*. It feels wide open in the moment, but it ignores repair costs, reversibility, and other people’s admissible sets. It tends to collapse after a few unreparable moves.
2. “Keepable” freedom – *anything I can keep doing*. This selects only from options that remain compensable given your current ledger (load) and horizon (time/energy/opportunity). It prefers reversible moves and early repairs, so that more options stay available later. It also works to widen others’ menus, so the social world stops fighting against you.

LoF argues that the second kind of freedom isn't *smaller* — it's *longer*. Like using good protection while climbing, keepable freedom lets you venture further because any falls won't end your journey.

20.1.2 The bill of keepable rights

These aren't moral commandments — they're mechanical principles that help preserve your ability to choose. Consider writing them down somewhere visible:

1. The right to reversible first steps. *Draft* before you broadcast; *test* before you deploy; *sample* before you commit.
2. The right to postpone irreversible harm. If something can hurt and cannot be undone, let at least one night pass before you decide.
3. The right to close human loops early. Say thank you, apologize, clarify, confirm — repairs made today buy you options tomorrow.
4. The right to bodily resets. You are allowed to eat, drink water, take medicine, get some light or darkness, go for a walk, or take a nap. Resetting your body widens your future menu.
5. The right to ask for specific help. “Can you check the second paragraph?” beats a vague “Could you help me?” Specific asks can expand both your menu and someone else’s.

6. The right to name your real horizon. Plans that ignore your actual time and energy are counterfeit freedoms.
7. The right to protect others' admissible sets. Don't "win" today by bankrupting someone else's tomorrow — you'll only inherit that debt later.

Keep these rights in mind as filter questions when you face a big choice.

20.1.3 The three-R test (a 10-second check)

Before you act, quickly ask yourself:

- Reversible? If this choice turns out wrong, can I back out of it with minimal cost?
- Repairing? Does this action reduce some outstanding social debt or task debt?
- Respectful? Will it leave others with viable options (i.e. it doesn't trap or harm anyone else)?

How to use it: If you get two "yes" answers, go ahead. If you get only one "yes," either scale the plan down or wait a bit. If zero, pick a different move entirely. This tiny rubric lets you implement QS-friendly decision-making without any math.

20.1.4 How guardrails feel from the inside

People often misinterpret QS's pruning as just mood swings, personal weakness, or "self-sabotage." Try reframing those experiences:

- The sticky no. You *could* send the cutting message, but your body hesitates and the words feel heavy. That friction isn't a flaw in you — it's your system protecting future compensability.
- The easy yes. Suddenly a nap, a walk, or an apology feels like the simplest thing to do. That sense of ease is your menu tilting toward needed repairs.
- The narrow day. When you're in pain, grieving, or up against a deadline, you have fewer keepable moves. That's physics, not personal failure. In those moments, pick just one "R" (Rest, Relief, Repair, or Realistic progress) and secure one small, true win.

20.1.5 Freedom inside guardrails is relational

Your choices shape the menus of those around you — family, colleagues, community. Two practical upgrades can foster mutual freedom:

- Offer menu wideners, not demands. “If I draft the outline, could you handle the slides?” widens the other person’s admissible set. In contrast, “I need these slides by 5 PM” might collapse it.
- Share horizons early. Try saying, “I have 40 good minutes of focus today — what’s the most helpful way I could use them for us?” This kind of transparency replaces blame with cooperative planning.

Communities that start talking in terms of menus, ledgers, and horizons build much more durable cooperation, because everyone is protecting each other’s future freedom.

20.1.6 Four freedom myths (and the reality)

1. Myth: “Freedom means *no* limits.” Reality: Freedom means having limits that keep doors open even after you walk through them.
2. Myth: “If I don’t seize this opportunity right now, it’s gone forever.” Reality: If something *must* be seized right now and is irreversible, it’s probably a trap. Let one night pass — if an opportunity can’t survive a pause, treat that urgency itself as part of the risk signal.
3. Myth: “Self-care is selfish.” Reality: Self-care is menu maintenance. Neglecting your own well-being today will shrink *everyone’s* options tomorrow.
4. Myth: “Setting boundaries will hurt our love.” Reality: Boundaries prevent unpayable debts, which *preserves* love rather than reducing it.

20.1.7 The keepability planner (5 minutes each day)

Once a day, try this simple planning exercise:

1. Name your horizon – realistically, how many hours of productive time and what energy level (low/medium/high) do you have today?
2. Pick one “R” as your daily theme – Rest, Relief, Repair, or Realistic progress.
3. List three reversible steps you could take today that align with that chosen R.
4. Close one loop – identify one person you need to thank, apologize to, or follow up with today, and do it.
5. Pre-decide one “no.” Write down one tempting but non-keepable action you will decline today.

Remember, you’re not “optimizing” life; you’re stabilizing your freedom. This is about keepability, not maximal productivity.

20.1.8 When you must choose between values

Sometimes every option in front of you causes pain: for example, telling the truth might strain a bond, but staying silent might erode your integrity. In LoF terms, you’re choosing among *compensable harms*. Triage those hard choices like this:

- Favor the option that preserves the most future options (the one that’s more reversible).
- Favor the option that has a clear repair pathway (one where you *know* how you could apologize or make amends afterward).
- Favor the option that minimizes harm to others first, even at some cost to you now — because social debts compound the fastest.

Whatever you choose, immediately sketch out a repair plan: who you’ll need to reach out to, when, and how you’ll do it. Writing that plan is half the battle of making the repair.

20.1.9 Freedom across seasons

Our lives have seasons, and LoF can guide you through each one:

- Build season (wide horizons, light ledger): Stretch yourself, learn new things, invest in projects, and take bold *but reversible* risks.
- Repair season (narrow horizons, heavy ledger): Close loops, rest aggressively, stick to conservative policies, and avoid any new brittle commitments.
- Transition season (major changes like a new job, a new baby, or a loss): Declare a temporary minimum menu for yourself (e.g. “two essential duties + one R per day”). In transitions, success is measured in keepability, not speed or splendor.

Seasons change. Simply recognizing which season you’re in prevents a lot of self-judgment and planning mistakes.

20.1.10 A note on dignity and agency

LoF never excuses “*the system made me do it.*” Within the admissible set, you still have real choices. You can choose generosity over cynicism, truth over spin, patience over panic. Guardrails explain why some options *feel* available or not, but they do not excuse cruelty or selfishness. The mark of mature agency is to choose the most humane keepable option you can find — and then work to widen the menu for yourself *and* others tomorrow.

20.1.11 Where we go next:

Now that we've seen how personal freedom can persist within LoF's constraints, we'll turn to the question of meaning. In the next section, we explore "Meaning Without Illusions," asking whether life can feel purposeful without assuming any grand cosmic design – and how restoring balance in ourselves and others might become a source of genuine purpose.

20.2 Meaning Without Illusions

Thesis: You don't need cosmic favoritism, secret purposes, or guaranteed happy endings to live a profoundly meaningful life. Meaning emerges when felt experience is coherent, connected, and keepable *within* the natural constraints of LoF.

20.2.1 Three pillars of non-illusory meaning

1. Coherence — “*Why this, why now?*” Life feels more meaningful when today’s efforts fit into a story you can tell without any hand-waving. Coherence isn’t about prophecy or knowing the future; it’s about having a traceable link between your actions and your values (“I showed up for my sister today because family repair keeps our future open”). LoF actually helps here because it favors repairs and reversibility — moves that make the narrative of your life hold together.
2. Connection — “*Who benefits with me?*” We experience meaning when our choices expand other people’s menus — their options for relief, repair, and growth. This isn’t performative self-sacrifice; it’s mechanics. Widening someone else’s admissible set widens the shared future for both of you.
3. Keepability — “*Can I carry this?*” Any sense of meaning that demands endless self-harm or fiction will eventually collapse. Choose commitments you can *sustain* given your ledger and horizon — often smaller, but truer, and longer-lasting. When it comes to meaning, keepability beats grandiosity.

20.2.2 How meaning is made (without magic)

- Values → Habits → Traces. Pick a small set of core values (say, *care, craft, truth*). Turn them into daily habits that leave tangible traces — thank-you notes written, code cleaned up, misunderstandings repaired. Over time, those traces accumulate into a life you can point to without hedging.
- Repairs are plot twists that redeem. On a hard day, the quickest route to a feeling of meaning is to make a repair. Close a loop, apologize, tidy up a neglected corner of your world. A repair doesn’t rewrite the past; it re-balances a future you can believe in.
- Witness beats audience. You don’t need applause or awards. One honest witness — a colleague, partner, or friend — who can say, “I saw you do the careful, kind thing,” will anchor your sense of meaning more firmly than a hundred internet “likes.”

20.2.3 Five practical moves to deepen meaning right now

1. Name the value, the person, and the step. For example: “Value = honesty; Person = Maya; Step = send the correction by noon.” Big, abstract values become meaningful when you attach them to a specific person and a concrete action.
2. Close one loop before you open a new one. Starting something new is exciting, but it drains meaning if old promises are left to rot. Aim for one closure per day — that’s compound interest for significance.
3. Make reversibility your studio. Use drafts, pilots, and prototypes so you can chase real ambition without wrecking keepability. Meaning needs a runway, not a cliff.
4. Trade spectacle for craft. Do at least one thing carefully enough that it would pass a blind review. Quiet excellence will impress your future self far more than any showy quick win.
5. Write the receipt. Keep a one-line daily log of the repairs and care you delivered: e.g., “Called Dad (repair), submitted preregistration (truth), stretched (body).” Seeing this evidence accumulate stabilizes your sense of meaning when your mood is all over the place.

20.2.4 Common illusions — and keepable replacements

- Illusion: “My life will be meaningful once I finally get recognized.” Keepable replacement: Recognition is just a side effect. Meaning is the sum of kept commitments that widen options — recorded in your own traces, not in other people’s headlines.
- Illusion: “Suffering automatically deepens meaning.” Keepable replacement: Only *repaired* suffering deepens meaning. Pain without repair is just pain; pain with repair can become binding strength.
- Illusion: “I must dedicate myself to one grand Purpose.” Keepable replacement: A portfolio of smaller purposes (e.g. care for X, build Y, learn Z), each scaled to your real horizon, is sturdier. Portfolios survive life’s weather better than any single, grand mission.
- Illusion: “If I feel uncertain, I must lack meaning.” Keepable replacement: Uncertainty is just the price of truth. Meaning grows by acting well in the midst of not knowing, not by pretending you’re certain about everything.

20.2.5 Meaning in hard seasons

When your horizon contracts (because of grief, illness, burnout, etc.), shrink your definition of a “meaningful day” down to one *R*:

- Rest – restore your body so that tomorrow actually comes.
- Relief – reduce the acute burden you’re under (even slightly).
- Repair – close one human loop or finish one lingering task.
- Realistic progress – take one reversible step on a valued path.

This isn’t “lowering the bar” — it’s choosing a standard that keeps life keepable. Your meaning doesn’t disappear in hard times; it just condenses into smaller, denser acts.

20.2.6 Community templates for meaning (without illusions)

- The Repair Hour. Once a week, everyone in the group takes an hour to close one personal or work-related loop and then shares a simple “Who/What/Done” report. Collective repair multiplies options and boosts morale.
- The Witness Swap. Pair up with a buddy and keep a two-minute voice-note log each day of each other’s kept commitments. Being *accurately seen* by someone is a powerful meaning amplifier.
- The Reversible-First Challenge. For 30 days, a team agrees to advance big ideas *only* through reversible moves. At the end of the month, they review how much progress survived without any heroics.

20.2.7 Meaning and mortality

LoF’s prediction of near-neutral closure at the end of life doesn’t trivialize the chapters of your life — it actually honors them. If every person’s ledger truly ends up roughly balanced, then the difference *you* make is in *how* the path was walked: how much repair you enacted along the way, how many options you preserved for others, how many truthful traces you left behind. Neutral closure doesn’t erase those traces; it just prevents life from turning into a worst-case moral lottery. In the end, your meaning is the pattern of care you kept: locally permanent in the lives it touched, and maybe historically permanent in the practices it started.

20.2.8 A pocket ritual (2 minutes, every night)

Each night, jot down a one-line answer to three questions:

- Truth: What’s one thing I faced honestly today?

- Care: Whose menu did I widen, even a little, today?
- Keepability: What did I do that I can keep doing tomorrow?

Write one sentence for each. By week’s end, you’ll have a handful of sentences that matter — read those six or seven lines back to yourself.

Meaning without illusions is essentially *evidence-based hope*. You don’t need guarantees from the universe — only a way to keep choosing reparative, reversible, respectful moves that hold together over time. That’s enough to build a life you could defend under cross-examination — and enough to help you sleep well at night.

20.2.9 Where we go next:

Having considered meaning, we next examine outcomes under each hypothesis. The following section, “If the Law Is True,” imagines what life and policy could look like if LoF holds: how we might respond with cautious optimism and align our caregiving or social strategies with nature’s balancing act.

20.3 If the Law Is True

Thesis: If the Law of Fairness *is* true — even if our measurements can only approximate it — then the most practical question isn't metaphysical ("Why is there suffering?") but operational: *How do we live well inside a world with guardrails?* If LoF holds, our priorities, institutions, and daily decisions should shift from "maximize the peaks" toward "keep options open, repair early, reduce uncompensable harms, and widen one another's menus." Fairness isn't something we plead for at the end; it's something we co-produce all along by making compensable futures more probable.

20.3.1 What changes for a person — *today*

1. You start optimizing for *keepability*. You choose goals you can keep pursuing even in rough weather. That means favoring reversible steps, regular repairs, and a sustainable pace. Peaks and triumphs are fine; it's the unrepairable valleys that the system prunes against.
2. You privilege repairs over indulgences. When two options give equal immediate benefit, you pick the one that closes a loop (apologize, clean up, finish a neglected task) rather than a fleeting treat. Under LoF, repairs carry a hidden "shadow price" bonus — they open up future options and stabilize the whole system.
3. You measure meaning by traces, not hype. You keep tangible evidence of care and truth — finished drafts, reconciled accounts, patient notes, fixed bugs. These traces outlast momentary mood swings and they mark your ledger's progress toward neutral.
4. You manage your horizons on purpose. When time feels short (end of a semester, a health scare, a goodbye approaching), you narrow your commitments. When time feels long, you experiment with reversible bets. This mirrors QS's natural menu tilt and helps you avoid costly stalls or regrets.
5. You pay attention to others' menus. You actively look for ways to widen someone else's admissible set — offering relief, access, clarity — because in a networked world, their freedom feeds back into yours. (Fairness is *networked*.)

(*Daily pocket checklist: try to include at least one act of Rest, one act of Repair, and one Reversible step in each day.*)

20.3.2 Relationships and teams

- Make early repair the default. In relationships or teamwork, small misalignments are cheap to fix if acknowledged quickly. The longer you ignore a problem, the more “interest” it accumulates and the heavier the future cost.
- Design for reversibility together. Use prototypes, trial periods, and soft launches in group projects or decisions — these are the interpersonal equivalents of reversible moves, and they preserve trust as well as compensability.
- Deliberately rotate horizons. Healthy teams cycle between “wide-horizon build sprints” and “short-horizon repair sprints.” In other words, alternate periods of ambitious expansion with periods of catching up and fixing things. That rhythm matches how admissible menus should breathe over time.

(Team ritual – 10 minutes/week: each member shares “one thing I broke, one thing I repaired, and one reversible next step” in our work.)

20.3.3 Health, therapy, and coaching

- Treat relief as infrastructure. Sleep, pain control, nutrition, and safe social contact aren’t luxuries; they’re the foundation that keeps the reparative set of options open. (It’s hard to make great choices while sleep-deprived or in severe pain.)
- Widen channels as the goal of care. In clinical and coaching settings, think of success as restoring the person’s access to options with high compensability. For example: providing effective analgesia so that a patient has the option to reconcile with family, or arranging transport so someone can attend therapy.
- Reframe the self-talk. Help people replace “What would the perfect me do?” with “What choice would keep tomorrow most compensable?” — The latter question outruns perfectionism by focusing on preserve-and-repair.

20.3.4 Ethics and justice

- Justice aims at restoration, not spectacle. If neutrality-at-closure is a real law, then justice systems should emphasize restitution, rehabilitation, and relational repair over pure retribution or public spectacle.
- Policy test: Ask of any policy, “*Does this reduce uncompensable harm and increase feasible repairs for the affected people?*” If not, it may just raise future shadow prices (future hidden costs) without addressing root causes.

- Ethical storytelling: Avoid narratives that glamorize taking irreparable risks or that frame suffering as proof of virtue. Under LoF, unrepairable harm isn't some currency that buys you glory — it's simply a dangerous breach of the constraints.

20.3.5 Education and culture

- Teach “reversible first steps.” Schools and training programs can reward drafting, iterating, and quiet, blind-reviewed craftsmanship over one-shot high-stakes performances.
- Normalize ledger literacy. Encourage students and community members to notice and track their small repairs and contributions the way they might track grades or achievements — because those small acts of care and fix-up are what keep a life keepable in the long run.

20.3.6 AI, product, and system design

- Guardrail-first design. In technology and product development, treat user well-being, reversibility, and graceful rollback as first-class objectives. For example, build *repair pathways* into every workflow — easy undo buttons, ways to appeal or correct mistakes, even tools to apologize or compensate for errors.
- Menu-widening moderation. In content platforms or social systems, don’t just suppress harmful content; also actively promote content and connections that *expand* users’ future options (teach a skill, provide support, build community) and that lighten uncompensable loads rather than adding to them.

20.3.7 Science and forecasting

If LoF is true, we expect several telltale empirical patterns (many of these were outlined in earlier chapters):

1. Horizon scaling: As a person’s perceived time remaining shrinks, their admissible menu *reliably* narrows toward more reparative actions. We also expect to see stronger inhibitory control kicking in against options with low compensability (risky indulgences, vindictive acts) as horizons shrink.
2. Repair premium: When immediate payoffs are equal, choices that involve repair (e.g. making amends or helping someone) should show an extra *value boost* in the brain (say, in vmPFC activity) and higher selection rates in behavior — especially when a person’s ledger is deeply negative.
3. Dream counterweights: After especially high-load days, the emotional content of dreams should skew in a compensatory direction (essentially “valence

inversion”). We’d expect more relief-oriented or processing themes in those dreams, and see that people behave more calmly or optimistically the next day (a kind of overnight easing).

4. End-of-life variance compression: In hospice or late-life trajectories, if channels are kept open (adequate pain relief, sleep, social contact), we should see moment-to-moment mood swings compress and drift toward neutral as death nears, with a corresponding uptick in repair-oriented behaviors (making amends, saying goodbyes).
5. Telemetry tilt: In real-world data, as known end-points approach (graduation, a project deadline, a terminal diagnosis), people’s “menus” should measurably tighten. We’d expect to see spikes in repair and closure behaviors and fewer wild, irreversible gambles during those endgame periods.

Policy implication: We should fund multi-site, preregistered studies to test these signatures across different labs and cultures. Emphasize ethical, low-intrusion data collection (telemetry) and collaborations with clinicians, so we advance the science *humanely*.

20.3.8 Worldview without superstition

- No teleology sneaked in. LoF doesn’t promise any cosmic intention or benevolence. It simply describes a constraint on which patterns of experience are sustainable for creatures like us. No hidden “purpose of the universe” here — think of it more like a conservation rule in physics than a moral plan.
- Hope with teeth. If LoF holds, you can rationally expect that even the worst moral “lotteries” in life are bounded — no one’s ledger goes infinitely negative without end. That hope isn’t for complacency; it translates into taking calmer, earlier action to repair things. (It’s hope that *motivates*, not hope that excuses.)

20.3.9 What becomes easier if LoF is true — and what stays hard

Easier:

- Letting go of status theatrics, because craftsmanship and honest repair would clearly yield the real long-term rewards.
- Making decisions under uncertainty, because you’d default to the choice that preserves future compensability (knowing the guardrails have your back on needless extremes).

- Talking about death, because “neutral closure” would give families a non-mystical, dignity-focused way to discuss the end-of-life chapter.

Still hard:

- Distributing scarce resources when many people’s ledgers are deeply negative. (*We’d still need fair institutions and tough ethical decisions.*)
- Telling genuine repairs from performative ones. (*We’d still need “blinding” and audits — i.e. objective checks — to keep us honest about what’s truly reparative.*)
- Living with irreversibility when channels are blocked. (*We’d still need human solidarity and creative innovation to handle situations where no easy repair is possible.*)

20.3.10 What changes *this year, and this decade*

This year (personal and organizational):

- Start a “Repair Hour” tradition and adopt a *Reversible-First* policy in your team or household decisions.
- Keep a simple private ledger of your care/repair/relief actions (just for yourself, to build awareness).
- In any product or policy you influence, introduce at least one built-in *repair pathway* (a way for people to undo mistakes, appeal decisions, get things fixed).

This decade (societal and scientific):

- Launch multi-site, preregistered studies on horizon effects and humane end-of-life dynamics — essentially, put LoF’s key predictions through the gauntlet with high standards.
- Pilot justice reforms that put repair first (e.g. diversion programs with restitution and “channel widening” support, rather than purely punitive sentences).
- Develop public tools for reversible commitments — think of civic “undo” buttons or platforms that make it easy to issue transparent corrections and second chances.

20.3.11 How to talk about it with people you love

- Plain language: “*I’m trying to choose the options that keep tomorrow fixable for us.*”

- Invitation: “*What’s one repair I could do that would make next week easier for you?*”
- Boundary: “*I’m not going to take steps that make our future uncompensable — even if they seem thrilling today.*”

20.3.12 What failure would look like (even if the law is true)

Even if LoF is a real principle, that doesn’t mean life can’t go horribly wrong. LoF-consistent lives can still fail by:

- Neglecting the channels. If pain, poverty, or isolation choke off a person’s options, their menu will shrink brutally. We still have to build and maintain those support channels.
- Neutrality ≠ nihilism. Neutral closure isn’t saying “nothing matters”; it’s saying “*how you travel is the difference you can make.*” Losing sight of that can lead to apathy or despair.
- Heroic overreach. Trying to repair *everything* all at once just creates new debts and chaos. Staged, steady repairs win out over grand, impossible gestures.

20.3.13 A two-minute closing exercise

Before bed tonight, jot down three quick lines:

1. One *repair* you will complete tomorrow that will make a future you care about more compensable.
2. One *reversible step* you’ll take toward a project that matters to you.
3. One *channel* you will widen (for yourself or someone else); for example, get better sleep, improve access to a resource, bring clarity to a confusion, provide transport, or manage pain relief.

Put that note where you’ll see it in the morning. Do this for a week and notice how your choices start to feel easier — not because fate magically favors you, but because your menu is finally aligning with the guardrails.

20.3.14 Where we go next:

We’ve explored the hopeful side; next comes the flip side. In “If the Law Is False,” we will confront the scenario that there is no automatic balance at work. Rather than despair, we’ll discuss why this possibility would *increase* our responsibility to actively relieve suffering, since nothing else will even the scales for us.

20.4 If the Law Is False

Thesis: Now suppose the Law of Fairness (LoF) is *false* — i.e. there really is no intrinsic constraint that guarantees each conscious life's net felt experience closes neutral by the end. What then? We need to be clear about the scientific, ethical, and personal consequences — and what remains worth doing regardless.

20.4.1 What “false” would mean (operationally)

LoF would be falsified if rigorous, preregistered tests repeatedly showed that:

1. No horizon scaling. As people's subjective horizons shrink, their menus *do not* systematically tilt toward relief/reparative actions beyond what ordinary fatigue, risk-aversion, or learning effects would predict.
2. No repair premium. After controlling for utility, conflict, arousal, habits, etc., choices to “repair” (make amends, relieve someone's pain) show no residual value boost in the brain or behavior, and agents following low-fairness (low-Φ) strategies do not reliably stall out.
3. No variance compression at end-of-life. Carefully blinded hospice studies find no drift of affect toward neutral, and no compression of mood swings, as death approaches — at least not beyond what medication or normal adaptation explains.
4. Dreams don't counterweight. After high-load days, people's dreams *don't* show any compensatory affect shifts, nor do their next-day behaviors ease in the predicted “rebound” ways.
5. No cross-context invariance. Our composite hedonic measures (like HCl) fail basic invariance tests (configural/metric/scalar) across cultures or age groups, even after careful calibration — meaning we can't even measure a “ledger” consistently across different people.
6. Rivals win cleanly. Other theories (predictive coding, reinforcement learning + homeostasis, etc.) can replicate all of LoF's supposed signatures *without* invoking neutrality-at-death — and those rival models out-predict LoF on fresh data.

If all of the above turned out to be true across multiple studies and populations, then LoF (as a law) would be considered *failed*.

20.4.2 Scientific consequences

If LoF doesn't hold as a law, researchers would likely conclude that:

- We're dealing with a tendency or local norm, not a fundamental law. The ups and downs of feelings would be explained by known processes like adaptation, allostasis (stability through change), learning effects, and culture-specific scripts — not by any global “path constraint” forcing a neutral sum.
- The Queue System (QS) would reduce to ordinary cognitive control mechanisms. In other words, all that talk of QS would just map onto well-known brain circuits for control and valuation (in frontal cortex, etc.), with *no special shadow-price* for compensability at play.
- The ledger is just bookkeeping, not a boundary condition. A life's ledger could still be a useful summary of experience, but we wouldn't expect it to gravitate toward zero by some necessity. Long-run totals of happiness or suffering might drift based on circumstances or personal policy, without any fixed endpoint.

20.4.3 Ethical consequences (what does *not* change)

Even if LoF were false:

- Relief and repair still matter. Pain control, rest, reconciliation, access to basic needs — these remain intrinsically valuable and *instrumentally* good for health, trust, and learning. We never needed a cosmic law to know that helping people and reducing suffering are good things.
- Duties of care remain absolute. Clinicians, caregivers, and institutions would still have the same obligations grounded in human dignity, rights, and harm reduction. Nothing about LoF's failure would license neglect or cruelty.
- Restorative justice still works. Repairing harm can outperform punishing harm on important outcomes and help restore function in society. Those practical benefits of restitution and rehabilitation don't rely on any neutrality-at-death principle.

20.4.4 Personal consequences

If there is no guarantee of a neutral ending:

- You hedge more against unrecoverable risks. Without an invisible guardrail, worst-case scenarios matter *more*, not less. You'd have even stronger reasons to choose reversible moves and make early repairs, because you can't count on “luck evening out” in the end.
- You find meaning without metaphysics. You'd ground your purpose in the tangible traces you leave — the value you create, the promises you keep, the suffering you alleviate — rather than in the hope of some final cosmic balancing.

- Hope, revised. Hope would shift from “the universe will even the score” to “we can build buffers, widen channels, and protect each other.” In other words, hope becomes a call to collective action and prudent planning, not an expectation of automatic fairness.

20.4.5 What the research program still gives us

A negative verdict on LoF wouldn’t make all the work pointless. It would still have delivered:

- Better measures (HCI). In developing LoF, we created composite, blinded measures of subjective experience (like the Hedonic Composite Index) that are extremely valuable for medicine, psychology, and policy in their own right. Making happiness and suffering measurable in a rigorous way is a win for science, even if the specific hypothesis fails.
- Horizon tools. Experiments that manipulate people’s sense of horizon (short vs. long term) are still very informative for understanding decision-making, willpower, regret, and risk management. Those insights are useful with or without a grand “Law.”
- Ethical telemetry. We built protocols for low-intrusion, consent-based monitoring of well-being (using wearables, apps, etc.) that respect privacy and autonomy. Those methods can set a new standard for humane human-subject research going forward.

20.4.6 The rivals we’d lean on

In place of LoF, we’d place more weight on other frameworks:

- Predictive coding / Free-energy principle: Perhaps affect is mainly tracking *surprise and uncertainty*. Comfort comes from reducing prediction errors, not from any fairness constraint. This theory can explain a lot about why we seek familiar, expected outcomes.
- Reinforcement learning + homeostasis: Maybe learning under biological drives (rewards and set-point adjustments) explains most of our return-to-baseline phenomena. Classic adaptation and opponent-process mechanisms might recreate a lot of what looked like “balance” in LoF, without needing a special law.
- Social scripts and institutions: It could be that norms, status dynamics, and cultural expectations shape the trajectories of our happiness more than any built-

in neural law. “Fairness” in life might come from social equilibria (or lack thereof), not from physics.

In practice, we would pivot to hybrid models that incorporate these ideas — for example, treating “making repairs” as one important learned strategy among many (not as a privileged path guaranteed to neutralize the ledger).

20.4.7 What we would retract or revise in this book

- Retract: Any claim that neutral closure is a *law of nature* or an inevitable outcome. We’d explicitly roll back any statements implying that everyone’s end-of-life feelings *must* converge near zero by some nomic necessity.
- Revise: The theoretical sections on QS would be reframed. We’d remove the idea of a special “shadow price of compensability” and instead explain any menu-narrowing effects with more conventional cost–benefit logic (changing horizons, fatigue, expected social returns, etc.).
- Keep: All the practical guidance for a keepable life — choosing reversible actions, repairing early, widening channels — because those behaviors robustly improve real outcomes even *without* a universal law enforcing them.

20.4.8 Where we go next:

Now that we’ve addressed both possibilities, it’s time to consider how to talk about them. The next section, “Talking with Skeptics,” will offer guidance on discussing the Law of Fairness with doubters. We’ll see how to frame LoF as a testable idea and invite critical collaboration, maintaining a tone of humility rather than dogma.

20.5 Talking with Skeptics

Aim of this section: You won't earn trust by speaking louder or leaning on vibes. You earn it by (i) stating the Law of Fairness clearly, (ii) explaining how it could be proven false, (iii) separating your ethical commitments from your evidence, and (iv) inviting tests that you might lose. This section offers a practical playbook for having constructive conversations about LoF.

20.5.1 Start by scoping the claim (what LoF is — and isn't)

- What LoF claims: Every unified conscious stream ends with a neutral ledger of felt experience at the death of mind. This is enforced not by magic but by global constraints that act as guardrails (not goals) on what choices are feasible — and these effects get stronger as one's subjective horizon shrinks.
- What LoF *doesn't* claim: It does *not* promise that events will feel fair or that lives will be just or pleasant. It's not karma, not fate, not cosmic justice, not "everything happens for a reason," and it doesn't involve any afterlife or moral scorekeeping. LoF only predicts certain statistical patterns in menus, choices, dreams, and end-of-life feelings when channels for compensation are available.

For example, you might say: "*We're proposing a constraint on the path of experience — not a grand purpose behind the universe.*"

20.5.2 Lead with falsifiability

Show exactly how the idea could be proven wrong:

- *Horizon test:* If, as time runs short, people's options don't skew toward relief and repair (controlling for fatigue, risk, etc.), LoF is wrong.
- *Repair premium:* If after accounting for other factors we find no extra brain/behavior "value boost" for choosing to repair, LoF is wrong.
- *End-of-life compression:* If careful, blinded hospice studies show no drift toward neutral mood when patients have open channels (good care), LoF is wrong.
- *Dream counterweights:* If people's dreams don't tend to flip or alleviate the emotional tone after very hard days, LoF is wrong.
- *Invariance:* If our composite happiness index (HCI) can't be calibrated to mean the same across different groups (i.e. fails configural→metric→scalar invariance), LoF is wrong.

You can say something like: “*We have five concrete, preregistered tests that could flat-out kill this hypothesis. If those tests fail across labs, we’ll retract the whole idea.*”

20.5.3 Steelman common objections (and answer them)

1. “*This is just hedonic adaptation with fancy language.*” Steelman: Regular adaptation and opponent processes already explain a lot of why people’s happiness returns to baseline. What’s new here? Reply: LoF predicts some patterns that plain adaptation doesn’t – for example, that as someone’s horizon shrinks, their menu tilts more toward repairs, and that “repair” choices carry a special premium even when they’re not immediately pleasurable. We’re specifically testing for those horizon-dependent effects ($\Phi \times H$ interactions and QS residual patterns). If we don’t find them, then it *is* just adaptation and we’ll concede that.
2. “*Predictive coding (free-energy theory) already covers this.*” Steelman: A brain trying to minimize surprise/uncertainty could explain affect and behavior without invoking any fairness law. Reply: Predictive coding is a powerful framework, but LoF’s signature would be *neutrality at life’s end* and the compression of affect in endgame scenarios — things that pure “surprise minimization” doesn’t necessarily predict. We’ll be directly testing whether LoF’s predictions overlap with or diverge from predictive-coding predictions. If those other models explain our results better or outpredict us on new data, we’ll bow out.
3. “*You’re sneaking in teleology — like the system wants to be fair.*” Steelman: Phrases like “the system balances” or “guardrails steer us” sound like you think nature has intentions or goals. Reply: We’re not assuming any purpose or goal, only constraints. Think of it like a conservation law in physics, not like Mother Nature wanting something. The math is about boundary conditions, not endpoints that someone *wants*. And we absolutely agree: if adding these constraints to our models doesn’t improve predictive power, we’ll drop them. No teleology for teleology’s sake.
4. “*It’s unfalsifiable — you can always blame ‘closed channels’ for failures.*” Steelman: If the data don’t show balancing, you might just say “oh, that person’s channels were blocked, so LoF couldn’t act.” That’s a loophole. That’s why we preregistered what counts as an open channel. We only analyze cases where key channels (pain relief, basic needs, communication) are in place. If even then we see no balancing patterns, we’ll outright call it against LoF. In other words, we’ve removed that loophole by defining it upfront.

5. “*Your composite metrics of happiness are too squishy.*” Steelman: Self-reports can lie, physiology is noisy — your measurements might just be garbage in, garbage out. Reply: That’s why we’re using a blinded composite of self-report, behavior, physiology, and neural signals. And we’re putting a huge emphasis on measurement invariance (making sure a “7 out of 10” means the same feeling for different people). If our measurement tools don’t hold up — if they’re biased or inconsistent — that will count *against* our hypothesis. We’re treating that as a possible failure mode, not sweeping it under the rug.

20.5.4 A respectful talk map (how to structure the conversation)

1. Clarify terms. For example: “*By ‘fairness’ here I mean the net felt experience closing neutral at the end — I don’t mean moral justice.*”
2. State the constraint clearly. e.g.: “*The idea is that certain guardrails prune down your options, especially when you feel you’re running out of time.*”
3. Acknowledge rival theories. Name other explanations you respect (adaptation, predictive coding, etc.) and the parts of experience they explain well.
4. Offer your kill-tests. Explain the key experiments that could invalidate LoF (those five from section B).
5. Invite collaboration or input. For instance: “*Would you design the test differently? Want to preregister something together?*”
6. Close with ethics. Emphasize: “*Regardless of how the tests turn out, we all agree on caring for people — relief and repair are non-negotiable.*”

20.5.5 Sample micro-dialogues

- Skeptic: “*Aren’t you just rebranding ordinary self-control?*” You: “They do overlap — brain regions like the rIFG and ACC are central in both. But LoF predicts something extra: as your time horizon shrinks, there’s an internal ‘shadow price’ that specifically boosts the appeal of reparative options even when the immediate payoff is the same. If we don’t see that effect (the $\Phi \times H$ interaction) after controlling for everything, then I’m wrong.”
- Skeptic: “*Give me one risky prediction this theory makes.*” You: “Sure. It predicts that at the end of life, if a person’s channels are kept open (good pain control, communication, etc.), their moment-to-moment feelings will compress toward neutral. Not necessarily turn happy — just narrow toward the middle. So if careful,

blinded hospice studies across different sites show no such compression, we retract that claim.”

- Skeptic: “Isn’t ‘neutral at death’ impossible to prove or disprove for an individual?” You: “Exactly — we don’t try to ‘judge’ any single person’s death. We look at patterns in groups. We set up criteria ahead of time and use blinded raters. We’re testing a statistical pattern, not giving each person a score.”

20.5.6 Etiquette with hard topics (suicide, trauma, injustice)

- Lead with care. Say something like: “*Nothing in this theory is meant to minimize anyone’s pain. We stand firmly against harm and for immediate relief whenever someone is suffering.*” Always make it clear that human well-being is the priority.
- No “loophole” talk. Never suggest that someone could “game” the system or that suffering now is okay because of some future payoff. For example, don’t ever imply suicidal thoughts are validated by LoF balancing things out. Emphasize getting real help, using crisis resources, and ensuring safety first.
- Separate science from support. You might say: “*Our study is asking whether these guardrails exist. But regardless of the answer, our duty to care for people remains exactly the same.*” This draws a clear line: the research question is separate from how we treat people in pain.

20.5.7 What to concede early (credibility boosters)

- Acknowledge that there are scenarios where rival theories might outperform LoF. “*For instance, in a quick reflex task with no emotional weight, standard reinforcement learning might predict behavior just fine without any LoF effects.*”
- Admit that if measurement invariance fails — meaning if our happiness metrics don’t work equally across different people — then cross-person ledger comparisons are off the table. We’d have to limit our claims to within-person patterns.
- Be clear that LoF is not a comfort doctrine. It doesn’t magically make life fair. Tragic outcomes can still happen. LoF isn’t about assuring anyone that “everything will be okay” — it’s a hypothesis about how systems behave, not a promise of cosmic justice.

20.5.8 Offer concrete collaborations

- *“Let’s co-design an experiment where we manipulate someone’s time horizon in a decision task, preregister everything, and split the data analysis. It’d be great to have you on board.”*
- *“Join our invariance working group—we’re actively trying to ‘break’ our own metric (HCI) across languages and age groups to see where it fails. Your perspective could help.”*
- *“Help us set up some adversarial tests. For example, define negative control scenarios or challenge LoF with a strong RL/predictive model and see where it wins. We’d love to collaborate on that.”*

20.5.9 A pocket checklist for public conversation

- Define your terms — e.g., explain exactly what you mean by “fairness,” “ledger,” “horizon,” and “admissible set” in simple language.
- Distinguish fairness vs. justice – make sure it’s understood that *fairness* here refers to an experiential balance, not moral or social justice.
- Disclose your kill-tests – up front, share the five key tests that could falsify LoF, and mention any negative controls you’re using.
- Defer to care – emphasize that regardless of the results, providing relief and opportunities for repair remains a moral imperative.
- Document and invite scrutiny – invite people to collaborate, preregister studies, or share data. Transparency goes a long way.
- Don’t—don’t promise any cosmic reward or consolation, don’t use teleological language (no “nature intends...”), and don’t speculate beyond what your data can actually support.

20.5.10 Where we go next:

Finally, after discussing dialogue, we’ll conclude the chapter by equipping you with a handy summary. The upcoming section, “A One-Page Summary to Share,” distills the key claim, evidence plan, and ethical stance of LoF onto a single page. It’s a resource you can use to quickly and responsibly communicate the core ideas to others.

20.6 A One-Page Summary to Share

Across a single conscious life, the total felt ledger must close exactly balanced at the moment of irreversible loss of conscious access (death of mind). The system's guardrails shape the admissible action menu so that no reachable trajectory can end with an unbalanced ledger.

Formal boundary condition: $L(T) = \int_0^T F(t) dt$, constrained to lie within a preregistered neutral margin around 0.

Observed ledger: $\hat{L}(t) = \int_0^t HCl(\tau) d\tau$, where HCl is the empirical proxy for $F(t)$ (in discrete data, replace the integral with a sum over samples weighted by the sampling interval).

Mechanism (notation): The Queue System yields $\mathcal{A}(t; \hat{L}, H, C)$, where C denotes available channels; the shadow price λ increases as horizon H shrinks; Φ denotes compensability features used to weight/prune options.

What LoF is (and isn't).

- Is: A constraint on experience (law-like, non-teleological), stated as a boundary condition at T.
- Is not: Karma, fate, or a moral scoreboard. It offers no license to harm and no excuse to delay relief.

How the guardrails work (plain language).

- At each moment, before you “choose,” your brain narrows the admissible menu to options compatible with closing the ledger at T. As time compresses, λ raises the repair premium and pushes toward relief, repair, and closure.
- Options that would make a neutral ending implausible lose salience or “stickiness,” while options that preserve compensability remain easy to access. (You still steer; the rails constrain feasibility, not authorship.)

How we measure the path (not the purpose).

- HCl (multi- channel composite) tracks momentary affect; integrating HCl gives the observable ledger $\hat{L}(t)$.
- We never infer cosmic intent; we test whether the boundary condition leaves measurable fingerprints in choices, sleep, and end-of-life dynamics.

Testable signatures (what must show up if the guarantee holds).

1. Horizon scaling: As H shrinks, admissible options narrow and tilt toward repair/relief. Operationally, menu counts should decrease with H^{-1} and exhibit overdispersion consistent with a Negative Binomial model (Poisson as baseline; NB if dispersion > 1).
2. Repair premium: When immediate utilities are tied, reparative options are chosen more often than non-reparative ones (with corresponding valuation/control-related signals, e.g., vmPFC and ACC, without assuming reverse inference).
3. Dream counterweights: After tough days, REM content shows valence inversion above chance levels (consistent with offline balancing passes).
4. End-of-life compression: With channels open (analgesia, sleep/dream, social contact), affect shows reduced variance and drift toward the preregistered neutrality band as T approaches, after adjusting for medication effects and reporting artifacts. This operational signature is a measurable consequence of the boundary condition; it is not the definition of the law itself.
5. Measurement invariance: HCI shows configural → metric (scalar where powered) invariance so comparisons are meaningful.

How we judge results (operational, not a retreat from the guarantee). These are testing margins, not the definition of the law. They exist because measurements are noisy:

- EOL neutrality/compression gates (preregistered): mean within ± 0.15 z; slope within ± 0.05 z/day; variance ratio ≤ 0.80 vs. matched baseline. Claims are gated behind these operational thresholds.
- Cross-group claims: require configural → metric (scalar where powered). If metric fails, we restrict to within-person claims.

Ethics we won't cross. "Relief is a systems variable; comfort and dignity override data collection." Balance at the end is never an excuse in the middle.

How to explain it in 15 seconds (guarantee wording).

The Law of Fairness says the felt ledger of a single life must close at zero at the death of mind. The brain's guardrails shape real options—especially as time runs short—so actions that relieve, repair, and close remain admissible and those that would leave a materially unbalanced end state are pruned.

We test this with HCI-based ledgers, horizon effects, dream counterweights, and end-of-life compression under preregistered neutrality gates.

20.6.1 Where we go next:

With the everyday implications and outreach tools now in place, we turn to Part X — Applying the Law: From Habits to Labs. First up is Chapter 21, “The Ledger Gym: Habits, Queue Traps and Repair,” where we translate the Law of Fairness into daily training: conditioning the Queue System through ordinary routines, spotting and escaping queue traps, and applying clean repair moves with simple ledger metrics. That practical footing then sets up the lab work that follows, carrying the same ideas into structured small-study methods and, in the next chapter, a full scientific playbook for testing.

Part X — Applying the Law: From Habits to Labs

Imagine two scenes: In one, a high school classroom is buzzing as students plan a small experiment to see if tough days really do lead to kinder dreams. In another, a team of seasoned researchers huddles in a lab, crafting a detailed protocol to test the Law of Fairness across multiple clinics. Part X bridges these worlds. Having laid out the theory, evidence, and ethics in previous parts, we now turn to practical experimentation — from simple personal trials to full-scale scientific studies. This part is a hands-on guide for everyone interested in testing LoF, whether you’re a curious student or a professional scientist. We’ll start with everyday life and small studies that anyone can try (in a classroom, at home, or in your own routine), then scale up to the gold standards of research design that experts demand. The goal is to empower you to engage with the Law of Fairness actively: to experiment, measure, and observe the effects for yourself, all while upholding the highest ethical standards.

In Chapter 21, we focus on personal and small-scale applications. We introduce the idea of the *Ledger Gym* — viewing your daily habits, emotional hurdles, and social interactions as exercises that keep your “life ledger” balanced. You’ll see how common practices (waking up early, exercising, fasting, sleeping well, practicing self-control) can act as informal tools to maintain emotional equilibrium. We’ll reinterpret age-old pitfalls (the so-called “seven sins”) as queue traps, purely in mechanistic terms, showing how extreme behaviors can throw off your internal balance and invite corrective backlash. We’ll also explore how our relationships and communities provide external *guardrails* — through apology, support, and even minor conflicts that help nudge us back on track. We connect these insights to longstanding cultural wisdom: many spiritual rituals (fasting, sabbath rest, confession, giving to others) can be seen as queue-management techniques that cultures may have developed to preserve well-being. To make it tangible, Chapter 21 invites you to conduct a safe N-of-1 experiment on yourself — a mini study to observe LoF dynamics in your own life. Finally, we close with essential warnings about what not to do with these ideas, underscoring that LoF is never a license for harm or complacency.

Chapter 22 then scales up from the personal to the professional. It lays out a Scientific Playbook for rigorously testing the Law of Fairness in the lab and the field. If LoF is true, it deserves the strongest possible evidence; if it’s false, we need to find out decisively. This chapter provides a blueprint for how researchers can design studies that *anyone* would trust. We’ll cover how to pre-register every hypothesis and analysis in detail (so nothing is left to bias or hindsight), how to build reproducible data pipelines (so results can be double-checked by independent teams), how to coordinate multi-site collaborations to

replicate findings, and even how to invite skeptics to poke holes (red-team challenges) to ensure the theory is tested from all angles. In short, Chapter 22 is a comprehensive toolkit for turning LoF from a bold idea into a *credible scientific program*. By the end of Part X, whether you’re a student running a class project or a scientist planning a large trial, you’ll have a clear roadmap to follow.

What this Part will do for you:

- Practical tools for testing LoF at any scale: Learn simple protocols that students and citizen-scientists can carry out (like tracking moods and dreams, or trying a short horizon decision game) as well as advanced methodologies for large studies.
- Ethical and effective study design: Understand how to introduce basic blinding and consent for small experiments (even with teens or classmates) and see the non-negotiable ethical standards for serious research (ensuring no harm and full transparency).
- A blueprint for rigorous science: Discover the elements of a “decision-grade” research plan — from locked preregistrations to multi-site replications — that will make tests of the Law of Fairness robust against bias and credible to skeptics.
- Bridging personal insight and professional inquiry: See how individual experiences (your own “ledger gym” practice) connect to the larger scientific effort. Part X shows that *anyone* can contribute to understanding LoF, and it sets the stage for the collective investigations to come.

Chapters in this Part:

- **Chapter 21 — The Ledger Gym: Habits, Queue Traps and Repair** - reinterprets everyday habits and social practices as exercises in maintaining emotional balance. It shows readers how to “train” their own Queue Systems through daily routines, avoid classic pitfalls that destabilize the ledger, and even run a small self-experiment. This chapter connects personal life tweaks with the Law of Fairness’s predictions, all while emphasizing safety and ethics.
- **Chapter 22 — The Scientific Playbook** - provides a thorough guide to formally testing LoF in research settings. It covers how to create bulletproof preregistrations, develop reproducible analyses and open-data practices, organize multi-site studies for stronger evidence, and encourage adversarial collaboration (challenge trials) to vet the theory. The chapter culminates in a one-

page “kit” that professional researchers can adopt immediately to begin credible LoF experiments.

Where we go next:

With the stage set, we now turn to Chapter 21, where we step into *the Ledger Gym*. In the coming chapter, we’ll explore how daily life itself can be approached like a training regimen for fairness — beginning with the simplest of routines and building up to deeper social and cultural practices.

Chapter 21 — The Ledger Gym: Habits, Queue Traps and Repair

Can life's ordinary routines and challenges function like a gymnasium for emotional balance? Chapter 21 answers yes — our daily habits, choices, and even setbacks can be seen as exercises that keep the “ledger” of our experiences in shape. This chapter demonstrates how seemingly mundane actions (like getting up early, exercising, or skipping that extra dessert) and social practices (like apologizing when you've hurt someone) serve as informal tools for the Law of Fairness in action. In other words, life itself offers a training program for our mind's fairness mechanisms. By viewing habits through a scientific, non-moralizing lens, we uncover how they may help our internal Queue System (QS) even out the highs and lows. We will also recast familiar human pitfalls — often moralized as the “seven deadly sins” — as queue traps, describing in mechanistic terms how these behaviors distort the balancing process and can lead to pain later. Crucially, nothing in this chapter preaches virtue for virtue's sake. Our approach is experimental: treat habits and behaviors as inputs to a system, observe the outputs (how they make you feel), and adjust accordingly, much like a scientist (or a gym trainer) would.

To keep things practical, we introduce simple metrics and protocols so you can test these ideas in your own life. By the end of the chapter, you should see everyday events — your morning alarm, your meals, a moment of anger or temptation — in a new, mechanistic light. “Living with the Law,” a concept introduced in Chapter 20, becomes concrete here: it means actively working with your brain's natural guardrails instead of against them. We begin with foundational daily habits, then progress to more complex “workouts” for your ledger: handling deeper psychological traps, leveraging relationships and community for balance, and tapping into age-old rituals for stability. A special N-of-1 experiment section invites you to be both scientist and subject, running a structured self-study to observe how changing a habit might shift your mood in line with LoF. We conclude with important warnings — drawing firm lines around ethical and safe practice, so no one twists the Ledger Gym idea into something harmful. Throughout, the tone is empowering and realistic: the Law of Fairness, if it holds true, isn't magic or karma, but it offers a hopeful framework. This chapter is about *living actively* with that framework — training wisely, avoiding known pitfalls, and respecting the limits — so that you can help life's balance play out with a little more grace.

What you'll get from this Chapter:

- A hands-on guide for applying the Law of Fairness in daily life. It introduces the metaphor of a “Ledger Gym” – viewing life's routines as workouts for your emotional balance. Each small action or challenge (waking up, choosing your

meals, resisting impulses) is explained as a training exercise for your brain's balancing mechanisms. Rather than prescribing morality, the chapter shows how habits function as tools that help your internal ledger stay even. This warm-up section prepares you to see all advice that follows through a mechanistic lens: your everyday life is the gym floor where your mind practices compensation and recovery.

- Common daily habits as simple guardrails or repairs for your emotional ledger. For example, waking up early provides a consistent emotional baseline; exercise delivers quick "repairs" to a low mood; fasting or moderation prevents extreme peaks; good sleep resets the ledger each night; and practicing self-control averts big negative spikes. Each habit is explained in scientific terms so you understand why it maintains balance, rather than labeled simply good or bad. These sections make it clear that a morning run or a healthy meal can have a measurable effect on mood stability – it's a practical input within the LoF framework, not a moral duty.
- Reframes familiar "sins" as seven queue traps – specific ways behavior can flood or starve the system. For example, overindulgence (gluttony) floods your system with pleasure and guarantees a crash later, while pride blocks the feedback that would keep you safe. Each of the seven traps is explained neutrally, highlighting cause and effect: why these behaviors often lead to imbalance. By understanding these traps in LoF terms, you gain practical insight into avoiding (or mitigating) those destructive cycles in your own life.
- Show how relationships act as external guardrails and repair kits. It explains that acts like apologizing and forgiving function as emotional repairs when we hurt someone, helping to clear interpersonal debt. Shared joy and support from friends or family can amplify your positives and buffer your negatives when you need it. Even minor frictions or conflicts serve a purpose – they are early warning signals that prompt you to correct course before a major rift forms. In this view, no one's ledger exists in isolation: friends, family, and community continually provide feedback and assistance, essentially forming a social Queue System that helps keep each individual within safe emotional bounds.
- Interpretations of rituals and cultural practices as formal queue hygiene. For instance, rituals like fasting after feasting, a Sabbath day of rest, confession or atonement, giving to charity, and daily meditation are shown to align with the Law's logic. Each practice is explained mechanistically: fasting counteracts indulgence, rest days prevent burnout, confession/forgiveness clear accumulating guilt, charity diffuses envy, and meditation resets the mind. Across cultures and history, such traditions have evolved because they naturally help

keep people's ledgers from spiraling out of control. Readers are invited to see the secular value in these old practices and to consider personal equivalents (like a "digital Sabbath" or journaling) to maintain balance.

- A step-by-step N-of-1 (single-person) experiment so you can be both scientist and subject. You'll learn how to pick one habit change, gather mood baseline data, implement the change, and then revert to measure its effect on your emotional ledger. This personal experiment makes the science real at your own scale. The chapter stresses ethical guidelines throughout, reminding you never to harm yourself or others as data. It closes with critical warnings: LoF is a descriptive principle, not a scheme to game the system. You are warned not to seek suffering in hopes of future reward and reminded that professional help and compassion are always preferable to fatalism. Overall, this chapter empowers you to live with the Law of Fairness: train your fairness system with good habits, learn from your social ties and rituals, avoid known traps, and use these ideas to foster empathy and patience – all while staying grounded in common sense and compassion.

Subsections in the Chapter

- **21.1 Warm-Up: Your Life as a Ledger Gym** – Introduces the Ledger Gym metaphor. This subsection explains how *everyday life* functions as a training ground for emotional balance under the Law of Fairness. It sets the stage by suggesting that each routine, challenge, or choice you face is like a "workout" for your mind's balancing mechanisms. Rather than moralizing habits as good or bad, we frame them as practical tools that help keep your internal ledger near neutral. By adopting an experimental mindset toward your daily activities, you can observe how different actions affect your feelings and learn to work *with* your brain's natural guardrails (instead of fighting against them). This warm-up prepares you to see all subsequent sections – from simple habits to complex social dynamics – in a new, mechanistic light.
- **21.2 Daily Habits as Guardrails and Repairs** – Surveys a range of common habits (like waking up early, exercising, practicing moderation in diet, prioritizing sleep, and exercising personal restraint) and shows how each serves as a guardrail or repair mechanism for your emotional well-being. In this subsection, we reinterpret these habits scientifically: early rising can provide a consistent emotional baseline, exercise can offer quick mood "repairs," fasting can help prevent extreme highs and crashes, good sleep can help reset the ledger nightly, and self-control can help avert actions that would cause big negative spikes. The focus is on understanding *why* these routines help maintain balance (through the

Queue System and ledger concepts) rather than treating them as virtuous in a moral sense.

- **21.3 Seven Queue Traps (Classical “Sins” Reimagined)** – Reframes the seven deadly sins (gluttony, greed, lust, sloth, wrath, envy, and pride) as queue traps – specific patterns of behavior that distort the normal balancing process of the QS. Each of the seven traps is explained in neutral, mechanistic terms: for example, overindulgence (“gluttony”) floods your system with pleasure and guarantees a crash later, while pride blocks the feedback that would keep you safe. By removing moral judgment, this subsection highlights the cause-and-effect nature of these behaviors: why they inevitably lead to pain or correction down the line. Understanding these traps in LoF terms gives you practical insight into avoiding them (or mitigating their damage) in your own life.
- **21.4 Social Relationships: External Guardrails and Repair Kits** – Explores how our interactions with others help keep our individual ledgers balanced. This subsection covers key social mechanisms like apology and forgiveness (which act as formal repair tools when we harm someone, releasing emotional “debt”), shared joy and support from friends or family (which amplify positives and buffer negatives when our own internal resources fall short), and the role of minor frictions or conflicts (small signs of irritation that serve as early warnings to correct behavior before a major rift occurs). We see that no person’s ledger exists in isolation – healthy relationships continually provide feedback and assistance that keep us within safe emotional bounds. In LoF terms, society itself functions as a larger Queue System, offering external guardrails and collaborative repair opportunities.
- **21.5 Ritual and Reflection: Spiritual Practices as Queue Hygiene** – Interprets various spiritual and cultural practices (such as fasting after feasting, taking a Sabbath day of rest, confessing and atoning for wrongs, giving to charity, and daily prayer or meditation) as strategies for queue hygiene. Without invoking any supernatural claims, this subsection shows how these rituals align with LoF’s logic: they regularly trim emotional extremes and foster balance. For instance, periods of fasting counteract indulgence, rest days prevent burnout, confession/forgiveness clear guilt before it accumulates, charity shifts focus from envy or greed, and meditation resets the mind. The idea is that over centuries, human cultures converged on these practices because they help keep individuals’ ledgers from spiraling out of control. Readers are invited to see the secular value in these traditions and perhaps adopt personal equivalents (like a “digital Sabbath” or journaling) to maintain their own balance.

- **21.6 N-of-1 Experiment: Training Your Own Guardrails** – Provides a step-by-step guide for conducting a one-person (*N-of-1*) study on yourself to observe LoF dynamics. This subsection empowers you to be both the scientist and the subject: you'll choose one habit change (like adding exercise or improving sleep), gather baseline data on your mood, implement the change for a couple of weeks, then see what happens when you revert to your old routine. It walks you through defining metrics (daily mood ratings, a simple Hedonic Composite Index, journaling notable events), structuring the experiment in phases (baseline “A,” intervention “B,” and return “A”), and analyzing the results. The goal isn’t to prove the Law of Fairness with one personal test, but to give you insight into how compensation or balance might manifest in your own experience. This hands-on section makes the science real at a personal scale, while stressing ethical guidelines (e.g. don’t harm yourself for data).
- **21.7 What Not to Do: Ethical and Practical Warnings** – Closes the chapter with critical “don’ts” to prevent misunderstanding or misuse of the Ledger Gym concepts. It clearly states that LoF is a descriptive principle, *not* a scheme to game the system. This subsection warns against deliberately seeking suffering in hopes of future reward (life doesn’t work like a bank where you deposit pain for interest), cautions that LoF never excuses hurting others (“they’ll get compensated later” is not acceptable), and advises against fatalistic or magical thinking (don’t attribute every event to cosmic balancing or avoid taking action because “LoF will fix it”). It also reminds you not to neglect professional help or personal goals under the guise of letting LoF run its course — you must still actively care for yourself and others. The final takeaway is about maintaining integrity and perspective: use these ideas to foster empathy, patience, and active coping, but always stay grounded in common sense and compassion. Life’s “ledger” is not a game to rig, and the true aim is a healthier, more balanced life, not some perfect score.

Where we go next:

We begin our exploration in 21.1, warming up with the idea that every *single day* of your life is effectively a training session in the Ledger Gym. Before diving into specific habits and techniques, let’s first reimagine our daily routine through the LoF lens.

21.1 Warm-Up: Your Life as a Ledger Gym

Every day, from the moment you wake, you are training under the Law of Fairness – whether you realize it or not. Just as a physical gym strengthens your muscles, life's routines and challenges exercise the “ledger-balancing” mechanisms that LoF hypothesizes in each of us. In other words, *simply living your life* puts your mind's fairness system through its paces.

Consider a typical morning: the alarm rings and you have a choice to get up or hit snooze. You might not think of this as anything important, but in LoF terms, even this small decision nudges your day's emotional ledger. Do you start with a sense of accomplishment and extra time (a small positive), or with stress from running late (a small negative)? Countless little moments like this accumulate. Every irritation at work, every kind act you do, every temptation you resist or give in to – they all register on your internal ledger of feelings.

The idea of a “Ledger Gym” is that these seemingly mundane habits and choices are essentially *workouts* for your Queue System (QS). They help your mind and body handle emotional ups and downs within natural guardrails, much as regular exercise helps your body handle physical stress. Each challenge or routine is training your system to compensate, recover, and stay balanced. Seen this way, your ordinary day is full of informal training sessions: a difficult conversation that builds your “resilience muscle,” or a moment of patience that stretches your “self-control muscle.” Importantly, everyone is enrolled in this gym automatically – there’s no opting out of life’s training. The question is whether you train *well* (using good form, so to speak) or let bad habits leave you off-balance.

21.1.2 Habits as tools, not moral imperatives

This section introduces how adopting certain daily habits (and avoiding certain pitfalls) can keep your internal ledger in shape. Crucially, none of this relies on moral judgment – we’re not prescribing virtue for virtue’s sake. Instead, we’re viewing habits as *tools*: practical levers that the QS might use (or that we can use consciously) to maintain balance. In the traditional view, habits like rising early or exercising often come with value judgments (“good” habits) and moral weight. Here, we strip that away. Eating a balanced meal or going for a run isn’t “good” because of some Puritan ethic – it’s useful because it can affect your brain’s ledger mechanics. Likewise, sleeping in or indulging in junk food isn’t “bad” in a moral sense, but it might carry a cost to how smoothly your mind can balance the books later. By reframing habits in this neutral, mechanistic way, we can talk about them without any shaming or preaching.

You'll notice throughout this chapter that behaviors often labeled as sins or virtues are reinterpreted in terms of cause and effect. A habit either helps your system stay within healthy bounds, or it tends to push you toward extremes that then require painful correction. We avoid loaded terms like "discipline" or "vice" here – instead, we ask, *what does this habit do for the ledger?* For example, does having a consistent bedtime prevent emotional volatility the next day? Does skipping a meal once in a while actually make it easier for your mood to stay steady? By treating habits as experimental tools, we empower ourselves to use them pragmatically. You're not waking up early to be "righteous" – you might do it because it leaves you feeling more balanced. This perspective also means there's room to forgive slip-ups (no moral failing, just a tweak to consider) and to personalize your "workout plan" for balance. Different tools may work better for different people, and that's okay. What matters is the outcome: a more stable, neutral ledger in the long run.

21.1.3 An experimental mindset for daily living

Consider it an *experimental* mindset for living: you're going to treat your day-to-day activities as inputs to a system, watch how they affect your feelings, and adjust accordingly. In practice, this means approaching your own life a bit like a scientist (or a curious athlete in training). Try viewing your choices and reactions as things you can modify and measure. For instance, if you suspect that scrolling on your phone late at night makes you groggier and moodier tomorrow, treat that as a hypothesis. Change the input (cut off screen time earlier for a week) and observe the output (do you feel calmer or more refreshed next day?). This chapter will equip you with the concepts to do this systematically by the time we reach 21.6, but the attitude starts now.

By the end of this chapter, you'll see your morning alarm, your meals, your social interactions – even your moments of temptation or anger – in a new mechanistic light. Mundane experiences transform into data points: oversleeping isn't "sloth" or "laziness," it might be a sign your system needed recovery (or that staying up late incurred a debt); feeling a pang of guilt after snapping at a friend isn't just "conscience" in the abstract, it's your QS flagging an imbalance that might need repair. You'll understand how "living with the Law" (as introduced in Chapter 20) can mean *living actively*: working with your brain's fairness guardrails, not against them. Instead of fighting your emotions or berating yourself for feeling down, you'll learn to ask: what counterweight might my system be seeking? Rather than seeing life's ups and downs as purely circumstantial or fated, you begin to recognize patterns of compensation and balance at play – and you take part in guiding them. And just like any good gym session, we'll start with basics and build up to more complex routines. First, we'll get the fundamentals of daily habit "exercises" down,

then move on to heavier lifts like overcoming major queue traps and leveraging community support. By approaching this progressively, you can develop a feel for the process before tackling the harder challenges. Ultimately, the Ledger Gym is about proactive engagement: you can't control everything life throws at you, but you can choose to respond in ways that nudge your ledger toward balance. That choice – made day after day – is the heart of training under LoF.

21.1.4 Where we go next:

Now that we've set the stage by viewing life itself as a kind of training gym, we turn to the *specific exercises*. In 21.2, we examine everyday habits – like waking up early, exercising, eating and sleeping wisely, and holding our temper – to see how they function as natural guardrails and repair routines for our emotional ledger.

21.2 Daily Habits as Guardrails and Repairs

Many of the habits people swear by – early wake-ups, exercise, periodic fasting, regular sleep, and general self-restraint – turn out to function as guardrails for our affective lives. Under LoF’s lens, these habits are not about ascetic virtue; they’re mechanisms that keep your hedonic ledger from veering to dangerous extremes. In plain terms, a good habit helps prevent an emotional blowout by providing steady maintenance. These behaviors prevent backlog in the “queue” of experiences by providing frequent, small repairs. Think of each positive habit as tightening a loose bolt before the machine falls apart. By regularly engaging in these practices, you avoid accumulating so much unchecked negativity or overstimulation that a huge correction becomes necessary. Let’s examine a few such habits and how they map onto the QS model:

21.2.1 Early rising and morning routine

Waking up early (and consistently) imposes a gentle discipline that resets your internal state each day. Mechanistically, a regular morning routine reduces chaotic swings in mood – it’s like starting each day with a calibrated baseline. From a Queue System perspective, an early, quiet morning provides low-conflict, low-arousal time to process any residual emotions from yesterday, effectively clearing minor negative entries from the ledger before the new day fully begins. Many people report feeling more “centered” on days they rise early – LoF would describe this as starting the day with Φ -positive choices (i.e. choices that have a high potential to be compensated smoothly if needed). For example, using the early hours to do something mildly effortful but rewarding – say, stretching, writing in a journal, or taking a brisk walk – adds a small positive entry to your ledger, compensating any overnight doldrums or anxiety. What seems like a matter of “personal character” (being a “morning person”) may actually reflect how your brain’s *menu* of options shifts when you honor your circadian guardrails. (Recall from Chapter 5’s sidebar “Menus vs. Personalities” that our so-called personality traits often align with the actions our QS has reinforced as safe or beneficial. An unreliable routine can feel like a character flaw, but it may simply be a sign that your internal menu isn’t getting the stable inputs it needs – early rising helps provide that stability.) In short, a regular morning habit keeps the ledger from starting in the red and primes you for balanced choices throughout the day. By beginning each day on steady footing, you’re less likely to make desperate moves later to correct a wobbly start.

21.2.2 Movement and exercise

Physical activity is a well-known mood regulator – a quick jog or even a walk can lift a negative mood and bleed off stress. In LoF terms, exercise is a classic *high- Φ action*: it has a high feasibility-of-compensation value (meaning it’s very effective at balancing out

prior negatives). When you've had a string of bad hours (or a bad day), engaging your body often produces a compensatory positive swing (endorphins, neurotransmitter resets, a psychological boost) that helps offset the prior negatives. It's no coincidence that we instinctively say "I need to walk this off" to clear our head after a frustrating workday – this is your internal ledger seeking balance. QS would predict that after sustained stress (when your ledger has drifted negative), the admissible set $A(t)$ tilts toward actions like exercise because they improve the odds of ending the day neutral. Indeed, if we could formalize it: $\Phi(\text{exercise})$ would be highest when stress is high, meaning the QS would flag exercise as an attractive option (even if you don't feel excited about it at first, there's often a gut-level pull to "do something active" when you're upset). The beauty of movement is that it's low-cost and immediate – a 20-minute walk can produce a genuine mood uptick, a small "deposit" of positive ledger balance that counters earlier withdrawals. Over time, habitual exercise builds not just physical fitness but emotional resilience. Your baseline mood tends to improve, and you recover faster from negative events. From the LoF perspective, you've trained your system to utilize a healthy repair strategy. Instead of stewing in frustration (which would leave a heavy negative entry queued up), you convert that energy into movement, which dissipates the negative affect and often brings perspective (small positives like a sense of accomplishment or relief). This isn't magical thinking or mere distraction; it's a direct physiological route to balancing the ledger (via endorphin release, reduced cortisol, and other neurobiological effects). It is well-established that even brief exercise can improve short-term mood and perceived stress. In summary, making exercise a habit means you've always got a *built-in repair crew* on call for your emotional ledger — and the more you use it, the more adept your system becomes at spontaneous self-repair.

21.2.3 Fasting and temperance

Deliberate restraint in consumption – whether skipping a meal occasionally (intermittent fasting) or moderating pleasures like alcohol, sweets, or screen time – might seem unpleasant for no reason. But through LoF's eyes, these practices serve as preventative maintenance on your hedonic ledger. Overindulgence in any pleasure can push your ledger far positive in the short term, only to invite a compensatory negative soon after (the "crash" after a sugar high, the malaise after a TV binge, the hangover after a party). By fasting or practicing temperance, you constrain the amplitude of those swings. It's a bit like applying a mild brake so you don't slam into a wall later. Mechanistically, short periods of discomfort (hunger pangs, a craving denied) might increase your subsequent sensitivity to pleasure in a healthy way – food tastes better after a fast, entertainment feels more rewarding when it's earned, etc. – thereby yielding a net neutral experience without the overshoot-and-crash cycle. In QS terms, choosing to abstain sometimes

keeps more options open later: you don't exhaust the "credit" of enjoyment all at once. There's also some evidence that intermittent fasting can improve mood and cognitive clarity for some; anecdotally, people often report a kind of calm or heightened focus on fast days. LoF would interpret that as the system adjusting: with a slight deficit induced (a controlled negative ledger entry), your body and mind may release counter-regulatory positives (e.g. increased alertness, a small sense of achievement). In effect, what looks like self-denial externally is internally a way to "stay in the green" by avoiding gluttonous spikes that demand painful correction. We avoid labeling this morally – it's not about "sinning" by eating cake or enjoying pleasures – it's about the physics of the ledger. A small hungry evening now might save you a week of sluggishness and regret later. Of course, moderation is key: the aim is gentle balance, not extreme deprivation (which would become its own trap).

21.2.4 Sleep and circadian rhythm

Nightly sleep is perhaps the most crucial built-in ledger reset – a natural *repair cycle*. Good sleep isn't just rest; it's an active balancing period when the brain processes the day's experiences. Dreams, as discussed in Chapter 10, often act as counterweights: after a day of heavy negative emotion, dreams tend to be emotionally lighter or even positive (and vice versa), as if to invert or soften the blow. In practical terms, maintaining a regular sleep schedule and ensuring adequate sleep length gives LoF's hypothesized mechanisms time to work. Think of each day as a mini-life: you accumulate positive and negative "ledger entries" while awake, and sleep is the end-of-day balancing where that running total is nudged back toward zero (through memory consolidation, hormonal resets, and the therapeutic effect of dreaming). When you shortchange sleep, you effectively carry an imbalanced ledger into the next day – small problems feel bigger, and your mood control is shaky. (A tired brain has a harder time generating the counter-regulatory responses that QS would normally deploy to keep you level.) Ever notice how after a string of poor sleep, you start getting emotionally "off-center"? Little annoyances loom larger, motivation drops (nothing feels quite as rewarding), and cravings for quick fixes (sugary foods, caffeine jolts, impulse buys) spike. In LoF terms, that can be read as a sign that λ (the shadow price on relief) is elevated: the system is starved for compensation, so it flags all sorts of easy-pleasure options as tempting, and your time horizon shrinks to just "get through the day." By contrast, after a solid night's sleep, people often report a *clean slate* feeling – even if yesterday was difficult, today doesn't feel so bad. This aligns with LoF expectations: sufficient REM and deep sleep may help process some of yesterday's pain (lowering the cumulative negative affect). **Repair and Sleep:** If you make one "ledger gym" investment, make it sleep. Regular sleep hygiene (consistent bed and wake times, a wind-down ritual, etc.) is like giving your QS a nightly

workshop to mend the ledger. For instance, following a hard day (say you received bad news or fought with a friend), your brain during sleep will often produce dreams that symbolically reverse the theme (Chapter 10 discussed evidence consistent with this inversion effect). You might dream of reconciliation or of succeeding at something you failed during the day – a subconscious attempt to provide emotional counterweight. You wake up not with the problem erased, but with its sharp edges sanded down, feeling more capable of coping. Recognizing this, you can treat going to sleep upset as a *queue hygiene* issue: unresolved extreme feelings at day's end are like an overfull queue going into the night, risking nightmares or restless sleep. It's no wonder there's folk wisdom advising "Never go to bed angry." LoF's take: if something is weighing your ledger down heavily at night, either resolve a bit of it (talk it out, write in a journal, pray or meditate – whatever provides partial relief) or be prepared for your brain to work overtime in dreams to compensate. In summary, sleep is your built-in ledger repair crew. Honor it – because if you do, it will reliably honor the balance for you.

21.2.5 Personal restraint and emotional regulation

Beyond scheduled habits, there's a more general skill of restraint – not sending the cruel text when angry, not hitting "Buy Now" on every impulse, not doomscrolling to infinity when you feel lonely. This kind of self-control can be seen as respecting the *invariance gates* your mind tries to set. Earlier we argued (Chapter 4) that LoF acts as a constraint, not a moral purpose – a guardrail, not a steering wheel. Exercising restraint is how you stay within those guardrails. For example, when you're furious (your ledger swung negative), there's a temptation to lash out because it might feel momentarily cathartic. But the QS "knows" – and indeed you've probably learned from experience – that hurting someone or burning a bridge will boomerang back with even more pain and regret later. The admissible set $A(t)$ in that heated moment *should* exclude, say, sending a hateful message or making a drastic decision. That option is effectively off-limits if your internal guardrail holds. When you actually hold your tongue or stay your hand, you are yielding to the guardrail and preserving future options (you can always express yourself later in a calmer way; importantly, you won't have to deal with the fallout of an impulsive blow-up). People sometimes call this *willpower* or *prudence*, but it might be better thought of as listening to the QS. Your body often gives cues – a tight gut, a flush of heat in your face – that an action you're considering is "high risk" for your emotional ledger. Restraint in that moment avoids creating a large negative entry (guilt, a damaged relationship, etc.) that would demand heavy compensation down the line. In the long run, cultivating this habit of pausing and *not* choosing the low-Φ action makes life smoother. It's akin to not yanking the steering wheel of a car – you prevent wild swings that require drastic counter-steering. Importantly, this isn't about suppressing your feelings for appearance's sake or adhering

to abstract virtue; it's about minimizing irreversibles. A harsh word can't be unsaid (it creates an enduring ledger debit until reparative work is done), but a thought unspoken causes far less lasting damage. By avoiding queue distortions in the first place, you need fewer heroic balancing acts later. Personal restraint thus aligns perfectly with LoF's ethos: let the natural balancing happen with as little violence (to yourself or others) as possible. Every time you resist an extreme reaction, you are essentially *preventing* a large future imbalance that you (or someone else) would have to suffer through and repair. It's a quiet habit, often unnoticed, but it is one of the most powerful for maintaining a fair experience over time.

21.2.6 Menus vs. personalities (sidebar)

Is it really “you” who loves sunrise jogs and quiet nights in, or is it your Queue System training you to keep your ledger in balance? This provocative question highlights a theme from Chapter 5 – our internal “menu” of attractive actions at any moment is shaped by past outcomes. Habits that keep yielding net-positive outcomes (or reliably neutralize negatives) get reinforced and start to feel like “my preference” or part of “my personality.” For instance, someone might say, “I’m just not a nightlife person; I prefer early mornings with tea.” On the surface that sounds like a personal identity statement. But dig deeper: perhaps for that person, late-night partying in the past led to emotional hangovers or anxiety (steep negative ledger swings), while calm morning routines gave reliably gentle positives. Over time, their QS narrowed the menu to favor mornings. Now it genuinely feels like a stable trait of their character. This perspective doesn’t diminish their agency – in fact, it reflects *wise self-regulation* honed by experience. Recognizing this interplay can be empowering: if you’ve labeled yourself “lazy” or “undisciplined,” it may be that your menu of available actions has been flooded with low-Φ temptations (junk food, mindless media) and your guardrails were obscured by repeated imbalances. By consciously tweaking habits (as we’re doing in this chapter), you’re effectively *retraining* your QS to prefer actions that make you – the experiential *you* – better off. In sum, what we often call *personality* can partly be the residue of ledger dynamics: a set of habits etched into us by what has historically kept us balanced (or imbalanced). LoF invites you to see those habits not as fixed character traits, but as *tunable parameters* in your life’s experiment. If some aspect of “who you are” is causing persistent pain or instability, it might not be “just who I am” – it might be a habit loop you can change. The Ledger Gym approach suggests that by adjusting your routines and listening to the feedback, you can gradually shift even those deep-seated patterns, effectively rewriting a bit of what feels like your personality in service of a more balanced life.

21.2.7 Where we go next:

We've seen how everyday habits can act as our personal trainers in the Ledger Gym. Now we turn to a different angle: what about the *bad habits* and extreme behaviors that throw us off balance? In 21.3, we'll explore seven classic "queue traps" — patterns of excess or avoidance that have long been warned against (think of the seven deadly sins). We'll strip away the old moral language and explain, in pure Queue System terms, why these traps reliably lead to suffering and how understanding them can help us avoid a world of hurt.

21.3 Seven Queue Traps (Classical “Sins” Reimagined)

Across cultures and history, people have noticed recurring patterns of human behavior that lead to pain and pleasure. Religious traditions famously codified some of these as the “seven deadly sins.” Here, we’ll describe seven queue traps – not as moral failings, but as distortions of the Queue System and ledger dynamics. Each trap is a way that the normal balancing process can be short-circuited or thrown off-kilter, often yielding a whiplash of compensation later. By understanding these in LoF terms, we can avoid their costs without resorting to moralizing. The idea is to see clearly the *mechanism* of harm in each case: how each of these patterns effectively fights against the Law of Fairness and thereby guarantees a rough correction down the line.

21.3.1 Overindulgence (gluttony)

This trap is all about ignoring the off-switch on pleasure. Biologically, when you’ve had enough of something good (food, drink, entertainment), your body sends signals – satiety, sensory fatigue – essentially telling QS, “Further consumption will yield diminishing or negative returns.” Gluttony is when we override those signals (perhaps due to abundance or engineered stimuli like ultra-sweet or ultra-addictive foods) and keep indulging regardless. The ledger distortion here is that you pile up excess positive entries in a short time – more than your system can smoothly compensate for. The result is a rebound negative: sickness, lethargy, or intense guilt once the pleasure fades. In queue terms, overindulgence *floods* the queue with one type of reward, causing a processing bottleneck – you become desensitized, and the later items in the queue (like basic needs for nutritional balance or simple bodily comfort) get delayed or denied. Eventually, the guardrail slams back in place: you might crash with a stomach ache, a pounding hangover, a wave of self-loathing, or chronic illness decades later that can provide years of consistent suffering. Importantly, this is not some mystical punishment for a sin; it’s a natural consequence of saturating a hedonic channel. The practical advice isn’t “don’t be greedy because it’s sinful,” but rather “overindulgence guarantees an equally extreme correction – save yourself that pain by enjoying things in moderation.” LoF predicts that keeping pleasures within reasonable bounds avoids provoking harsh compensations (no severe sugar crash if you never created a massive sugar high to begin with). In short, if you take too much now, expect to pay for it later – not by moral decree, but by homeostatic necessity.

21.3.2 Insatiability (greed)

Greed is a distortion where the ledger’s zero point is forgotten. The greedy individual keeps accumulating – money, power, achievements – without ever feeling satisfied, as if any positive balance must immediately be chased further. Mechanistically, this

resembles a *baseline shift*: the person's expectation or "neutral point" drifts upwards as they gain, so they never register a profit emotionally. It's like a hedonic treadmill where the speed increases every time you run faster. In QS terms, greed can be seen as a persistent refusal to let the ledger ever tip negative, combined with an inability to truly feel the positive (because it's hoarded or taken for granted, not savored). The distortion here is that the queue never clears – every fulfilled desire is instantly replaced by two new desires waiting in line. Over a lifetime, greed often invites drastic correction: either external (market crashes, losses, social isolation) or internal (burnout, a crisis of meaning). Why? Because an ever-upward chase of positives without natural pauses is unsustainable – it violates the equilibrium constraint. Eventually something gives: the person takes an ill-fated risk or alienates others (incurring large negatives that finally balance the books). Again, it's not cosmic vengeance, just cause and effect. Greed also narrows one's admissible set $A(t)$ to *only* those actions that increase the hoard, which cuts out much of what makes life fulfilling (relationships, simple pleasures). In effect, greed imposes a self-made poverty: a shrinking of life that paradoxically leads to a sense of impoverishment despite abundance. The take-home: striving and ambition are fine, but if you find that *nothing is ever enough*, consider that you may be running an unwinnable race against your own ledger. Letting yourself feel "full" (content, grateful) at times is not complacency; it's maintaining contact with reality. LoF suggests that if you never allow satisfaction, you're guaranteeing that balance will be forced upon you through some crash. In other words, if you don't allow small contentments, life may impose a big correction.

21.3.3 Runaway desire (lust)

Here we refer not just to sexual lust, but any kind of compulsive craving for intense pleasure (including addictive behaviors). Lust in LoF terms is when the promise of a very positive experience shorts out the evaluative process. The normal guardrails that check "Is this a good idea?" are bypassed by a surge of want. Neurobiologically, think of it loosely as the brain's ventral striatum (reward-related circuitry) drowning out the prefrontal control centers – QS is temporarily hijacked by raw craving. The queue distortion is that one particular reward stimulus comes to dominate the menu, eclipsing all other needs or considerations. For example, someone might pursue a new romance or sexual encounter despite it jeopardizing their long-term relationship, or binge on a drug at the expense of health and responsibilities. In that moment, the immediate high is so over-weighted that future consequences (ledger hits) are ignored – effectively, it's as if λ (the shadow price on future suffering) were set to zero temporarily when it absolutely shouldn't be. The LoF consequence is predictable: a large negative ledger imbalance accrues out of sight and then crashes in later – heartbreak, guilt, withdrawal pains, lost

trust or opportunities. Lust isn't "evil"; it's imbalanced trading: taking too much short-term credit and eventually having to pay a hefty debt. The wisdom across cultures has been not prudishness per se, but *balance*: intense pleasures are among life's joys, but when chased recklessly, they distort valuations and invite disaster. QS normally prunes extremely self-destructive options, but lust adds a false halo to them, tricking the guardrail into silence. One strategy that emerges (and indeed is found in many traditions) is to institute cooling-off periods: for example, wait a day before acting on a sudden overpowering desire. This gives the ledger time to reveal the true cost. Often, the craving will recede enough for perspective – QS "wakes up" and you realize, *oh, that would have been a mistake*. In sum, lust-type traps teach us that the intensity of a want doesn't equal its true value to our ledger. By inserting pause and reflection, we let the normal evaluation process catch up and protect us from our own momentary impulses.

21.3.4 Neglect (Sloth)

Sloth is not just laziness; it's *avoidance* of needed effort or repair. In ledger terms, this trap is letting negative entries accumulate due to inaction. Everyone needs rest, but slothful behavior goes beyond healthy rest into neglect: not fixing small problems while they're still small, not engaging with life's tasks because it's easier to do nothing. The distortion here is a failure in the compensation process: normally, discomfort or dissatisfaction spurs us to act (pain is a call to change). Sloth dampens that response – the queue of "things to address" grows longer, but nothing is dequeued. Imagine a sink piling up with dirty dishes because you can't be bothered; eventually it overflows and becomes a nasty, more intractable problem. Emotionally, it's the same: skip confronting an issue long enough, and the suffering often multiplies. In QS terms, a healthy system assigns growing urgency (Φ rising) to neglected tasks or duties – they *should* start to bother you more over time. If one consistently suppresses that, or if depression/low energy stifles the normal response, the backlog can reach a critical point. Then a crisis hits: deadlines all at once, a relationship breaks down from long neglect, your health collapses because you avoided all exercise or check-ups. Sloth, therefore, isn't just a personal failing in a moral sense – it's a dangerous strategy of deferring all compensation to the future. The ledger will balance, but the longer it's left unbalanced, the more drastic the eventual correction (like an untreated infection that becomes life-threatening). The practical LoF tip: do a little now rather than a lot later. Small regular efforts (even if unpleasant in the moment) prevent the mountain of pain that inactivity breeds. It's emotionally much easier to maintain a habit or address concerns incrementally than to face the alarm-bell-level intervention once the queue is completely jammed. Recognize too that a lack of motivation can be self-reinforcing – the ledger stays negative and that dampens motivation further (a queue trap indeed). Breaking out often requires forcing a

small action to get some positive momentum (a principle known in therapy as “behavioral activation,” where doing one tiny productive thing can start an upward spiral). In essence, avoiding all effort now means more pain later – so when you feel that minor nag of discomfort about something you’re neglecting, try to take it as a friendly signal from QS to do a bit, before it becomes a crisis screaming for a lot.

21.3.5 Reactive rage (wrath)

Anger is a natural emotion, but wrath as a trap is anger *unleashed without constraint*, causing disproportionate harm. In LoF mechanics, wrath is akin to exploding your ledger’s contents onto others – it may relieve your internal pressure momentarily (a positive blip for the angry person, a feeling of release), but it dumps a huge negative into the environment (often onto loved ones or colleagues), which ultimately reflects back on your own ledger. Because LoF posits that experiences balance *within* a conscious stream, directing harm outward doesn’t actually eliminate it; it ensures that the aftermath (guilt, damaged relationships, retaliation from others) will weigh down *your* ledger later. Wrath thus usually guarantees a painful compensatory cycle: you hurt someone, they react (maybe not immediately, but trust erodes or they lash back eventually), and you end up with pain in your life as a result. The QS normally steps in when anger flares – that tight jaw or internal voice saying “Don’t do it” is an attempt to throttle a low- Φ action. Wrath is when we ignore those signs and act as if there’s no tomorrow. Indeed, a fit of rage often comes with a sense of narrowed time: “*I don’t care about the consequences!*” – essentially a zero-horizon mindset. The result, of course, is that you create consequences that you *will* care about later, once your horizon expands again and λ (your sensitivity to future suffering) goes back up.

The ancient idea of wrath as “deadly” can be translated into LoF terms: uncontrolled anger is deadly to optionality – it kills future good options. The advice grounded in fairness mechanics is the same as the age-old advice, but now we see why: every time you refrain from lashing out and instead find a calmer outlet or delay your response, you prevent a cascade of negative ledger entries that would inevitably follow. Techniques like taking a deep breath, counting to ten, or physically leaving the room when enraged are not about politeness or submissiveness; they are *micro-strategies to keep your queue from being flooded* with irreversible actions that will demand painful repair. In short, wrath doesn’t solve the problem, it *becomes* the problem – so the key is catching anger early and giving it a chance to cool before it does lasting damage.

21.3.6 Comparison and resentment (envy)

Envy is the trap of ledger *myopia* – obsessing over others’ apparent surplus and your own perceived deficit, instead of minding your *own* process. When you’re envious, you treat life like a zero-sum ledger between people (“their success somehow makes my position worse”). This is a distortion because LoF operates at the individual lifetime level, not by direct comparison across individuals. Envy skews your perception of your own admissible set $A(t)$: suddenly, options that were perfectly fine (your steady job, your comfortable home, your caring partner) feel inadequate because someone else has “more.” This often leads to self-sabotage or unnecessary risk-taking – you might forsake a stable path that was working for you to chase what someone else has, or conversely you might become so disheartened that you take no action at all, letting your ledger stagnate in negativity. The crucial damage envy does is it injects poison into your own well – the other person’s fortune doesn’t actually add a negative to your ledger, but your reaction to it can. In QS terms, envy introduces a false failure signal; it’s like an internal alarm screaming “We’re behind!” when in fact you might be on track in your own journey. This can shut down genuine opportunities (for instance, you might refuse help or friendship from those you envy, cutting off potentially positive experiences) and can even lead to malice (trying to undermine the other person, which creates moral and emotional debt for you). From a fairness-law standpoint, envy is particularly futile: *someone else’s highs will not prevent or ensure your highs and lows* – each life balances on its own. So stewing in envy doesn’t hasten your payoff; if anything, it distracts you from noticing and seizing your own compensatory positives. The antidote is refocusing on *your* ledger. Practicing gratitude (literally writing down or acknowledging what’s going well for you) is one way to counter envy’s distortion by making your own positives more salient. Another is to consciously turn envy into admiration or learning: if you feel that twinge of resentment, ask “*Is there something they’re doing that I can emulate in my own way?*” This shifts the mindset from zero-sum to positive-sum, where another’s success can inspire actions that improve your ledger rather than subtract from it. In essence, envy wastes your time and energy on the wrong ledger – someone else’s – when all that truly matters for your well-being is how your own ups and downs balance out.

21.3.7 Hubris and overconfidence (pride)

Pride in the toxic sense is losing humility about the ledger – believing the rules don’t apply to you or that you’ve “beaten” the system. In myths, this is the classic hubris that invites Nemesis (a downfall). Stripped of mythology, the LoF reading is: extreme pride leads you to indulge in pleasure, ignore feedback, and skip repairs. If you’re convinced you’re always right or invincible, you won’t apologize when you should, won’t course-correct

after mistakes, and might take on wildly risky bets assuming you'll come out on top. Essentially, pride disables the alarm that signals "Hey, you're veering off – fix this." A prideful leader might alienate their team through arrogance; minor issues fester because the leader can't acknowledge them, until a major collapse happens (employees quit en masse or a critical project fails). Or consider health: pride might make someone refuse to admit they're struggling ("I don't need help, I've got this") – delaying therapy or care until their condition worsens dangerously. Pride is a queue trap because it blocks compensatory information. QS works best when we're honest about weakness and pain – that's how it knows to deploy countermeasures or seek support. If pride prevents vulnerability or admitting error, you effectively forbid others (or your own rational mind) from helping to balance your ledger. LoF doesn't "smite the proud"; rather, the proud smite themselves by ignoring early warnings. The bigger someone inflates their balloon, the more forceful the pop when reality (ledger equilibrium) catches up. The practical wisdom mirrors the moral lesson but with a mechanistic justification: stay humble because an accurate sense of your ups and downs keeps you safe. Humility means accepting small corrections (criticism, setbacks) and making repairs (apologies, adjustments) before they become crises. In LoF language, pride raises the risk of irreversibility – some mistakes, if not admitted and addressed in time, become unfixable. Humility keeps the system flexible: you'll bend before you break.

Taken together, these seven distortions show how easily the fairness machinery can be derailed by extreme behaviors or mindsets. Each "sin" reflects a violation of LoF principles: refusing balance (greed, pride), chasing all positives now (gluttony, lust), deferring all effort (sloth), or ignoring that actions echo and must be repaired (wrath, envy). By reframing them as queue traps, we emphasize cause and effect over moral judgment. The message is empowering: these traps are avoidable, not because some authority forbids them, but because of practical understanding. If you treat your life like a system that must stay within bounds (not too high or too low for too long), ancient advice like "be moderate, be patient, be humble" starts to sound less like preachy rules and more like savvy operating instructions for a well-tuned life.

Where we go next:

Now that we've dissected internal habits and pitfalls, it's time to look outward. In 21.4, we examine how our relationships and communities act as an external Ledger Gym, providing support and feedback that keep us balanced. We'll see that no one trains alone – family, friends, and society at large supply crucial guardrails and repair kits when our self-regulation falters.

21.4 Social Relationships: External Guardrails and Repair Kits

No person is an island, and under LoF this is a very good thing. Our relationships – family, friends, colleagues, community – act as an external system of guardrails that help keep our individual ledgers balanced. How so? Other people provide feedback, support, and sometimes pushback that correct us when our self-regulation falters. In fact, one could say that society is a macro version of the Queue System, wherein each of us, through interpersonal interactions, helps others not drift too far off track. Of course, relationships can also introduce new stresses, but healthy ones net out as stabilizing forces. This subsection breaks down a few key social mechanisms in LoF terms.

21.4.1 Apology and forgiveness (restoration)

When you hurt someone or break a norm, you've introduced a negative imbalance – not only in your ledger (guilt, remorse) but in theirs (hurt, anger). The act of apology is a formal repair mechanism to address that. Think of an apology as an attempt to cancel out the negative entry with a positive action: you acknowledge the harm (validating the other person's pain) and often promise to make amends. This is more than just social decorum. LoF-wise, it's a high-Φ move: it significantly increases the likelihood that the emotional ledger between you and the other person can return to neutral. A sincere apology, followed by forgiveness from the other party, can rapidly relieve guilt and prevent lingering emotional debt on both sides. Psychologically, both the apologizer and the recipient often feel a weight lifted after forgiveness – that “weight” is precisely the ledger load being released. Many cultures ritualize this process: from saying “I'm sorry” with a bow or a hand on heart, to larger reconciliation ceremonies or days of atonement. Why such emphasis on apology across humanity? Because over centuries we learned that unrepaired breaches fester (the debt of hurt compounds interest, sometimes becoming feuds or long-term self-loathing), whereas timely apology and forgiveness resets the local ledger before it scars. In Queue System terms, offering a genuine apology re-opens the menu of interactions; it puts a relationship back into a state where positive exchanges are possible again (versus a cold-war stalemate of mutual avoidance). It's important to note that none of this requires believing in cosmic karma or divine forgiveness – it's interpersonal mechanics. An apology is effective because it directly addresses the emotional math: it compensates the person wronged (through acknowledgment, empathy, or restitution), which in turn allows them to let go of the grievance without feeling cheated. Restoration is the broader concept here: after a harm, some act of restoration (be it an apology, a compensatory gift or service, or even just the sincere expression of regret) is needed to repair the tear in the social fabric. LoF would predict – and indeed we observe – that relationships with open channels for apology and

forgiveness tend to heal and even strengthen over time, whereas those lacking such channels accumulate scarring (or break apart entirely). We reinforced this idea back in Chapter 19 on ethics: nothing about LoF ever excuses *skipping* restoration – if anything, it demands it. In practice, a timely apology is one of the most powerful tools you have to maintain fairness in the shared ledgers you have with others.

21.4.2 Shared joy and support

On the flip side of conflict, relationships also amplify positives and buffer negatives. When something wonderful happens to you, sharing it with friends (a celebratory phone call, a dinner party) doesn't dilute the joy – it often increases it, because their excitement adds to your positive ledger. And when you're in pain, loved ones provide comfort that softens the blow. A hug, a listening ear, a reassuring "I'm here for you" are not trivial gestures; they are direct injections of positive affect at a time when you're running a deficit. In QS terms, our close contacts often serve as "surrogate QS agents" – they intervene when our own internal compensations aren't enough. For example, consider how a partner or a good friend might sense you've had a bad day and suggest, "Let's do something fun tonight," or simply give you extra kindness and attention. That's essentially an external trigger for a compensatory action (doing something high-Φ like a relaxing walk together or enjoying a favorite hobby) at a moment when you might not have mustered it yourself. Sometimes we need that nudge because our own system is stuck in the negative. Similarly, when communities rally around someone in crisis – organizing fundraisers for a family in need, bringing meals to a sick neighbor, or just checking in regularly – they infuse tangible positives to offset hardship. LoF underscores that these supportive acts aren't just niceties; they are integral to how humans achieve balance. A person isolated from social support has to rely entirely on their internal resources to counterbalance hardship, which is much harder. Thus, cherishing and cultivating relationships is, in a sense, a rational strategy to ensure that when your ledger skews, there are people who will help tip it back toward center. In practice, this means don't hesitate to share both good news and bad news with people you trust. It's not burdening others – it's allowing the natural social balancing to occur. Joy shared is joy multiplied; pain shared is pain divided. From an evolutionary view, humans likely developed these tendencies because groups where members uplifted each other would outlast those where everyone struggled alone. In your personal life, this translates to a simple LoF-informed habit: make time for friends and family, celebrate their wins, comfort them in losses, and allow them to do the same for you. It keeps everyone's ledger more balanced than going it solo.

21.4.3 Minor frictions (irritation and social signals)

Not all guardrails in relationships are warm hugs; some come as frowns, scoldings, or moments of friction. Surprisingly, these little negative feedbacks are *healthy signals* in the economy of fairness. If you are being rude and you see your friend’s eyes roll, or you hear impatience creeping into your colleague’s voice, that stings a bit – and ideally, it checks you. That mild social pain is a cue: *queue adjustment needed!* In an ideal scenario, you notice the cue, apologize or change your behavior, and the irritation dissipates – the ledger impact stayed small because it was caught early.

Think of social irritation like a rumble strip on a highway: the brrrr sound and vibration when you drift out of your lane. It’s not pleasant, but it alerts you *before* a full crash (a furious fight or a lost friendship) occurs. Chapter 19 discussed how respecting others’ boundaries and dignity is non-negotiable; here we add that everyday irritations are part of that boundary-setting in practice. They serve a compensatory function: they provoke a small guilt or concern in you, which – if you act on it (say, by immediately adjusting your behavior or offering a quick “Sorry, you’re right”) – prevents a larger guilt later. From the perspective of LoF, feedback-rich relationships (where people can safely express annoyance, concern, or disappointment in small doses) are self-correcting. In contrast, relationships with silent resentment (no expression of irritation until it explodes) tend to have big ledger swings – long periods of superficial neutrality then a big rupture. So, counterintuitive as it sounds, a bit of complaining or bickering can be a sign of a healthy dynamic that’s constantly balancing out minor inequities. We should note: chronic, hostile conflict is different and harmful; we’re talking here about *proportionate* signals – the gentle nag, the sarcastic “Oh, finally you’re here” when you’re late – that indicate an expectation of repair, not malice. Embracing these signals (rather than being immediately defensive) is key. If someone you care about is irritated with you, it usually means they value the relationship enough to want it fixed rather than to walk away. Listen to that cue: it’s much easier to apologize for being late today than to rebuild trust after many such slights go unaddressed.

In summary, our social network provides both safety nets (empathy, help) and guardrails (critique, consequences). Both aspects keep our lived experience within bearable bounds, ideally guiding us toward a life that, at its end, feels resolved and fair – not because every moment was easy, but because every strain was met with some counter-strain, every mistake with a chance to mend. Society, in effect, runs a “distributed ledger” alongside our individual ones: norms, laws, reactions and forgiveness all play roles in ensuring that no one person’s extremes become the community’s burden unchecked. We will revisit this phenomenon on a larger evolutionary scale in Chapter 18 (how social

norms might emerge to enforce balance), but for day-to-day life, the takeaway is simple: saying sorry, giving thanks, checking in on friends, or even voicing a complaint – these are all part of maintaining queue hygiene in the human family. They are the subtle exercises and corrections of our collective Ledger Gym.

21.4.4 Where we go next:

We've looked at personal habits, internal traps, and now the role of community in keeping the balance. Next, in 21.5, we turn to an intriguing question: what about the *rituals and reflections* humanity has developed over millennia? We will see how spiritual practices – from fasting to forgiveness rituals – can be viewed as cultural strategies for ledger hygiene, and what we can learn from them in modern life.

21.5 Ritual and Reflection: Spiritual Practices as Queue Hygiene

Why do so many religions ask us to fast, rest on certain days, confess our misdeeds, or regularly practice forgiveness and charity? These practices long predate any scientific theory, yet they map uncannily well onto LoF's idea of maintaining balance. In this section, we interpret several spiritual or religious practices as guardrails and cleanup routines for our emotional queues – all without invoking any supernatural teleology. Whether or not one believes in the spiritual rationale, these rituals often have *concrete psychological benefits* that align with keeping one's ledger near neutral and one's QS functioning smoothly. It appears that human cultures, through centuries of trial and error, stumbled upon habits that help lives stay balanced. Let's look at a few examples:

21.5.1 Fasting after feasting

Many cultural calendars feature a period of restraint following a period of indulgence. For example, Lent comes after Carnival, and in Islam the daytime fasts of Ramadan come after nights of feasting. Mechanistically, this is scheduled variance compression – a deliberate counter-balancing. After a time of plenty or excess, a time of self-denial restores equilibrium. This isn't framed (in our analysis) as moral penance; it's an intuitive recognition that *constant* indulgence dulls the mind and body, whereas a step back refreshes them. Physiologically, fasting recalibrates hunger and insulin responses; psychologically, it renews appreciation for the basics (food tastes almost magically good after a fast!). LoF-wise, one could say these traditions enforce a pause in positive intake so that the running ledger doesn't spiral and require a crash. By voluntarily embracing a controlled negative (hunger, abstinence), people often report spiritual clarity or heightened empathy for the less fortunate. In our terms, that's a perspective reset – a reminder of what genuine need feels like, which in turn makes one's ordinary life feel relatively rich (a cognitive swing toward the positive once the fast is broken). These cyclical fasts thus act as guardrails against gluttony and insatiability: they rhythmically prevent the queue from overfilling with pleasure to the point of numbness. It's like regularly cleaning the palate of life. While some may see fasting purely as religious obedience, on a functional level it serves the ledger: it keeps highs and lows in check through an oscillation that ensures neither extreme dominates for too long.

21.5.2 Sabbath and periodic rest

A weekly day of rest (be it the Jewish Sabbath, the Christian Sunday, or any sacred rest day in various traditions) can be seen as a horizon reset. In Chapter 6 we discussed how a shrinking time horizon increases the urgency (λ) to set things right. A weekly rest day, by design, structures your horizon into a repeating week-long cycle, prompting you to put aside work and worry at least for that one day. It's an enforced queue flush – you stop

adding new tasks or pleasures and take stock instead. Secularly, we know that overwork without rest leads to diminishing returns and burnout (in other words, huge negative ledger corrections like illness or breakdowns eventually force a stop). The Sabbath principle – “Six days you shall labor, on the seventh you shall rest” – ensures that no matter how driven you are, you regularly step off the treadmill of constant striving. This keeps long-term goals from completely eclipsing immediate well-being. How is this queue hygiene? It prevents backlog of life’s neglected facets. On a rest day, one might reconnect with family, reflect on life, or simply catch up on sleep – all actions that rebalance areas of life that get out of whack during the work week. LoF would note that Sabbath observance historically carries promises of renewal: it’s said to “restore your soul” in religious texts. Remove the poetic phrasing, and we can interpret that as literally giving your emotional ledger a chance to heal routinely. Many people report higher well-being when they take regular days off for genuine rest or spiritual practice, and it can help prevent severe stress accumulation. The key isn’t a divine reward; it’s that by resting, you avoid running huge chronic deficits of social connection, sleep, or contemplation. In LoF terms, Sabbath is a guardrail against slothful neglect on one hand (because it’s purposeful rest, not just procrastination) and against greedy overreach on the other (because it limits relentless accumulation and work). It provides a structured rhythm to life that mirrors the homeostatic rhythms of the body. Today, even in largely secular lives, the idea of a “digital detox” day or a weekend with no emails serves a similar protective function. Regular rest is not just indulgence – it is maintenance.

21.5.3 Confession, atonement, and forgiveness rituals

As highlighted in Chapter 17, many religions have formal processes for acknowledging wrongs and seeking forgiveness. Examples include Catholic confession and penance, the Jewish High Holy Days of repentance (Yom Kippur), Islamic repentance (Tawbah) and expiation (kaffara), and Hindu pujas that include asking pardon from deities and people. These can be seen as systematic ledger audits and resets. Instead of letting guilt or grievances silently accumulate over a lifetime, these traditions encourage at least a yearly “clean-up.” You examine your conscience, speak your failures, perform some act to make amends (perhaps saying specific prayers, doing charity, or directly apologizing to those harmed), and symbolically wash the slate. The effect on participants is often profound relief and a renewed sense of agency – in LoF language, Φ (feasibility of compensation) spikes upward because suddenly a lot of previously “stuck” negatives (sins, regrets) are lifted. These rituals are typically low-cost in material terms (mostly words, symbolic acts) but high-impact in emotional terms, which is exactly what you want for an efficient repair process. They formalize what might otherwise be awkward or avoided: they *give you permission and structure to do the very things that restore balance*

– admit fault, seek forgiveness, grant forgiveness. The Jubilee tradition in the Torah went so far as to institute a periodic reset every 50 years, including freeing indentured servants – a societal reset that parallels an individual forgiving all who wronged them in their heart periodically. While such extreme practices aren't universally adopted in secular life, the concept holds: make whole what you can, regularly. A skeptic might say, “People just confess and then sin again – what's the point?” From a mechanistic view, the point isn't instant moral perfection; it's preventing irreversible buildup. It's like clearing clogged drains before they overflow – you might have to do it again later, but you avoid disaster by doing it routinely. These practices also often reduce stress and rumination; knowing you have a moral safety valve can keep anxiety at bay. Indeed, as we theorized in Chapter 17, after confession-like rituals one should see measurable drops in stress markers and intrusive negative thoughts, and a boost in forward-looking positive action. Empirically, many people report feeling “lighter” or “reborn” after such rites – a poetic way of saying their ledger got a reboot. In everyday terms, even outside religious context, one can adopt a similar habit: perhaps a Friday reflection where you acknowledge whom you hurt or what you regret that week and then decide to make a small amends or self-forgiveness gesture. The specific form matters less than the function: doing periodic maintenance on your conscience unloads hidden burdens and lets you move on unencumbered.

21.5.4 Almsgiving and charity

Many faiths strongly emphasize giving to the poor or donating a portion of one's earnings – tithes, zakat (alms in Islam), dāna in Buddhism and Hinduism, etc. On the surface, this seems like a moral duty unrelated to one's own emotional balance – it primarily benefits others. But many people report that giving can increase personal happiness and a sense of meaning. LoF insight: compassionate acts are typically Φ -positive. When you help someone in need, you often feel good yourself (a warm glow, a sense of meaning or connection). So regular charity can actually function as a preventative counterweight to self-centered stress. It keeps your perspective broader: your own problems may feel lighter after you've actively helped with someone else's more dire needs. And it deposits positive feelings of social contribution in your ledger. Religions effectively bake this into life as a habit – not only for societal good, but likely because people who give regularly tend not to be as consumed by envy or greed. It's hard to feel your life is meaningless when you regularly see the difference you make for others. Almsgiving, therefore, combats the traps of insatiability and envy by redirecting focus and creating a sense of abundance (“I have enough to share”). It's queue hygiene in the sense of cleaning out the stagnation of purely self-focused pursuits. When you give, you make emotional space for new experiences and reduce the risk of getting stuck in a loop of “I, me, mine” which often leads to discontent. Even outside religious context, people who volunteer or donate to

causes tend to report greater satisfaction. The act of giving is a small ledger donation to someone else *and* to yourself: it reminds you that positive actions are readily available and effective. In sum, charity aligns individual and collective ledgers in a beneficial way – it creates positive-sum feedback. No wonder it's a pillar of so many ethical systems.

21.5.5 Meditation and prayer

Apart from doctrine, many religious people engage in daily prayer or meditation, often in the morning or evening. These practices encourage reflection, gratitude, and release. In our model, think of them as daily mini-ledger audits. A morning prayer might include hopeful intentions and seeking guidance for the day (setting a positive mindset, which primes one's QS to notice high-Φ choices throughout the day). An evening prayer might include thanking for good events and asking forgiveness for wrongs (essentially doing a small emotional ledger reconciliation before sleep). Meditation similarly clears the mind of ruminative thoughts, which can be seen as *flushing minor negatives* and resetting arousal levels. Many people find that such practices reduce stress and improve emotional regulation. LoF suggests why: by regularly attending to one's inner state, acknowledging emotions, and then deliberately cultivating calm or trust that "things will be okay," you leverage a top-down influence on the QS. You remind yourself of a bigger picture (which can lengthen your perceived time horizon, thereby lowering λ if you've been too narrowly focused on immediate troubles), and you often cultivate compassion (which increases pro-social, high-Φ impulses). The non-teleological take: even if one doesn't believe a prayer is supernaturally heard, the act of formulating one's worries and hopes, and then releasing them, is psychologically beneficial. It externalizes burdens (much like writing in a journal can do) and reinforces positive norms ("Help me be kind, help me be patient" – effectively you're priming your own guardrails for the next day). Thus, prayer or meditation are internal maintenance routines – sharpening the tools (awareness, patience, acceptance) that keep the ledger balanced. Regular meditation, for example, is often associated with greater emotional resilience and less reactivity, which in LoF terms means fewer wild swings that need compensation. Many traditions also include elements of *gratitude* in these practices ("count your blessings"), which directly counteracts envy and amplifies existing positives on one's ledger. In secular form, even a short daily mindfulness session or writing three things you're grateful for can mimic these effects.

In sum, spiritual traditions, when viewed through the Law of Fairness lens, appear to be early human experiments in *affective engineering*. They set up regular intervals for queue maintenance (daily, weekly, yearly rituals) so that individuals and communities don't veer off into destructive extremes. Importantly, one need not believe that these practices were

originally designed with LoF in mind; rather, cultures may have evolved them because groups that had such practices *fared better* – they enjoyed more cohesion, individuals were more resilient, and there was less internal strife or collapse. It’s a form of convergent wisdom: different religions came up with fasting, sabbaths, confession, almsgiving, and so on, all addressing similar underlying needs for balance and compensation in human life. As Chapter 17 noted, it’s as if culture itself implements a macro-QS, ensuring “horizon management” and routine repair opportunities are woven into the fabric of life. For a reader of this book, whether you are religious or not, the takeaway is that these time-tested practices have secular analogs: a periodic digital detox can be your “sabbath,” a personal journaling or therapy session can be your “confession,” volunteering can be your “alms,” and so on. By instituting your own rituals of pause, reflection, giving, and atonement, you create guardrails in your life course. You don’t leave balance to the very end of life; you pursue ledger hygiene throughout – which ultimately might make the final balancing act less dramatic and more graceful.

21.5.6 Where we go next:

We’ve now explored maintaining balance through habits, avoiding traps, leaning on community, and even drawing on cultural wisdom. But how can *you* directly observe these balancing effects in action? In 21.6, we’ll design a personal experiment. We’ll show you how to run a safe, one-person study – essentially turning yourself into a test subject – to see whether deliberate changes in your routine lead to the kinds of mood shifts the Law of Fairness would predict.

21.6 N-of-1 Experiment: Training Your Own Guardrails

This section turns the ideas of the Ledger Gym into a personal experiment. How can you see the Law of Fairness operating (or not) in your own daily life? While a single-person trial can't prove a law, it can yield insight and help you tune your habits for better balance. We'll outline a safe, ethical N-of-1 protocol – meaning *you* are both the researcher and the participant. Think of it as running a small study on yourself, using the concepts of HCI, QS, λ , and Φ as your guide.

21.6.1 Goal

The goal of this experiment is to observe whether deliberate habit changes lead to measurable shifts in your mood/experience balance, consistent with LoF predictions. We also want to test if specific compensatory phenomena (like dream counterweights or short-horizon urgency spikes) show up in personal data. Essentially, you're looking for *any* patterns indicating your own ledger pushes toward neutrality when tipped. If LoF has any sway in your life, a structured self-observation might catch glimpses of it. And even if it doesn't, you'll learn something about what affects your mood. Importantly, another goal is self-improvement: by identifying a habit that potentially unbalances you and attempting a fix, you might actually feel better by the end. So this isn't just about data; it's about personal growth through a scientific lens.

21.6.2 Study design

We'll use an ABA' design over a span of about 4–6 weeks. "A" is a baseline period with your normal behavior. "B" is an intervention period where you introduce a new habit or routine targeting ledger balance. Then "A'" is a return-to-normal period to see if things revert (this also helps avoid attributing any changes merely to the passage of time or placebo effects). This classic design lets you compare before, during, and after. Here's a step-by-step breakdown of the plan:

- Choose Your Focus Habit (Intervention): Pick one habit from 21.2 that you're not already practicing rigorously, and that you suspect might influence your mood or energy. Good options include: getting up one hour earlier than usual and using that time constructively (early rising + morning routine), committing to 30 minutes of exercise daily (if you're currently sedentary), a dietary tweak like no refined sugar or a 16-hour fast twice a week (fasting/temperance), or a nightly sleep hygiene routine with a fixed bedtime and no screens 1 hour before bed (sleep improvement). You could also try a mindfulness routine if stress is your concern. Important: Choose something *feasible and safe*. Do not pick an extreme or potentially harmful change (e.g. a 3-day water-only fast or sleeping only 3 hours a

night would be inappropriate and dangerous). The idea is a gentle, positive change, not a stunt to test your limits.

- Define Metrics (HCI and More): Decide how you will measure your well-being and ledger status throughout the experiment. At minimum, use a simple daily mood score – for example, each evening, rate your overall day on a +5 (great) to -5 (terrible) scale. This can serve as your simplified Hedonic Composite Index (HCI) for the day (a rough composite of your feelings). You should also track relevant objective metrics if possible: e.g., hours of sleep, minutes of exercise, or even physiological data if you have wearables (like resting heart rate, sleep quality index, etc.). Additionally, keep a ledger journal: jot down brief notes on any notably positive or negative experiences each day and their rough impact (e.g., “Got compliment from boss +2; argued with spouse -3; went for a walk +1; felt lonely in evening -2”). This doesn’t need to be exhaustive, just highlights. The journal will help identify *qualitative* compensation patterns (like “Bad workday followed by really uplifting evening with friends” or “Felt so down that I impulsively ate cake,” etc., which are the kind of narratives that numbers alone might miss).
- Baseline Period (Week 1–2, “A”): For one to two weeks, live as you normally do, but record the metrics you decided on diligently every day. Don’t introduce any new habit yet – this is your observation period to see how things are without intervention. You are establishing your baseline mood variability and any natural QS phenomena that occur. Be as consistent as possible with the tracking: same time each day for mood ratings, keep the journal entries short but honest. During baseline, also note any instances of what might be LoF in action: Did a string of stressful days lead you to crave something unusual (comfort food, a night out, excessive sleep)? Did you have a surprisingly good day after a lousy one (maybe a spontaneous “rebound” day where things just felt better)? Are your dreams noticeably related to how your previous day went (jot a word or two about dream emotion in your journal if you recall it)? At the end of baseline, you’ll have, say, 7–14 days of data. Compute or observe your average daily mood and its variability. Maybe you find your mood oscillated from, say, +1 down to -2 typically, with an average around -0.5 (slightly on the negative side). Keep those numbers; they’ll be your baseline comparison.
- Intervention Period (Week 3–4, “B”): Now implement the habit change you chose, and stick to it consistently for at least two weeks (longer if you can). Continue measuring everything exactly as before. This period is where you’re *applying a guardrail proactively*. LoF prediction would be that beneficial habits should either

raise your average mood, shrink the swings (lower variance), or both. For example, if exercise is your intervention, perhaps you'll notice fewer severe “–3” bad days because the stress is getting bled off via workouts, or an overall shift upward in baseline mood by a point. If improving sleep is your goal, maybe your journal will show fewer impulsive snacking incidents late at night and a more even-keel morning mood. Be attentive also to queue trap occurrences: do you find it easier to avoid the traps from 21.3 when your habit is in place? For instance, on days you exercise, are you less prone to irritability (wrath) or overindulgence? On days you get good sleep, do feelings of envy or greed diminish a bit because you're simply more content? Also watch for horizon effects: if your habit is morning-focused, does it change how you feel about the day's “horizon” (perhaps you feel more optimistic about getting things done)? If it's evening-focused (like a bedtime routine), do you notice a change in late-day urgency or closure (maybe you start naturally doing a quick reflection each night because you're calmer)? Record any such observations in your journal.

- Return to Baseline (Week 5+, “A’’): After a solid period of habit practice, you have the option (for the sake of the experiment) to *stop* the new habit and go back to your old routine for another week or two. (If the habit was extremely beneficial and you truly don't want to stop, you can modify this step or shorten it – ethics and well-being trump experimental rigor here. But ideally, try a brief return to baseline behavior for the data.) The reason to revert is to see if any gains quickly fade or negatives return, which strengthens the case that the habit caused the changes. For instance, if during the exercise weeks your average mood was +1 and in the stop week it falls back to –0.5, that's suggestive that exercise was helping. Or if your sleep quality drops again when you quit the routine, and you notice your daytime feeling and patience worsen accordingly, that's valuable evidence. This A' phase also helps control for the possibility of placebo or “measurement effect” – sometimes just paying attention in baseline improves things. If the positive changes persist even after stopping, maybe something else shifted in you more permanently (or other factors are at play). Document whatever happens, even if it's messy or inconclusive.
- Analyze and Reflect: At the end, compile your data. Plot the daily mood scores over the weeks if you can (a simple line chart can be enlightening) – do they show a visible difference in the B period compared to the A periods? Calculate averages for baseline vs. intervention. See how the variability changed (did the highs and lows tighten up?). Look at your journal: did you see signs of compensatory behavior or queue effects? For example, perhaps you notice: “*In baseline, every*

time I had a bad day at work, I coped by binge-watching TV (which left me groggy the next day). During my intervention of proper sleep and a no-screens-before-bed rule, after bad workdays I went for a walk or read, and the next day I felt okay." That anecdotal pattern is huge – it's exactly a guardrail working (you replaced a low-Φ coping behavior with a higher-Φ one and saw better next-day outcomes). Or maybe: "*Dream journal: 3 out of 5 really bad days in baseline were followed by unusually positive dreams; during the intervention, I simply had fewer really bad days to begin with, and my dreams were more neutral – maybe because there was less extreme to counteract.*" Observations like that speak to LoF's mechanisms (the dream counterweight effect, etc.). There's no right or wrong outcome; the point is to see *dynamics*. If you find none (e.g., the habit made no noticeable difference, or your mood was all over the place regardless), that's a result too. It could mean LoF isn't manifesting strongly at that timescale, or the particular habit wasn't a key lever for you. Real self-science is often like that – messy and requiring tweaks or a different approach next time.

21.6.3 What to watch for (ethics and pitfalls)

During the experiment, remain ethical to yourself. This means: do not push a habit to dangerous lengths for the sake of data. For instance, no extreme fasting or deliberate sleep deprivation just to "see what happens" – we already know those queue traps are well-documented and not worth the risk. If you chose a fasting intervention, do it in a healthy way (e.g., skip dessert or have an earlier dinner, not starve for days). If you find your intervention stressing you out (say you're utterly exhausted from waking up an hour earlier and it's hurting more than helping), adjust or abort it. The goal is improvement and insight, *not* self-punishment. Also, keep an eye out for unexpected compensation. Sometimes when you impose one change, something else shifts to fill the gap. You might find, for instance, that when you started exercising daily, you began eating more without realizing (your body compensating for calories) – does that have mood effects? Note it. Or maybe quitting late-night video games made you initially more irritable (because you lost a coping mechanism), but later you started calling friends during that hour (finding a healthier compensation). These second-order effects are valuable to notice; they show QS in action, figuring out new equilibria when conditions change. In essence, be both scientist and subject, but also your own caregiver throughout. If something feels off or harmful, stop or modify it – life is longer than this little experiment, and your well-being comes first.

21.6.4 Possible outcomes

Ideally, you'll come away with a personal narrative of how a conscious habit tweak influenced your emotional balance. Maybe your average daily HCI (mood score) went up a bit (not dramatically, but noticeably). Maybe the lows weren't as low, or the highs didn't crash as hard. Or perhaps you found you had more energy (a positive ledger drift) and fewer cravings for queue-trap behaviors (like less urge to stress-eat or doomscroll). A successful outcome might read like: "Before, I had a rough day about 3 times a week and usually coped by binge-watching TV (which left me groggy and guilty). During my experiment where I enforced a consistent sleep schedule and wrote in a gratitude journal at night, rough days dropped to maybe once a week, and I woke up feeling less anxious overall. When I stopped those habits in A', the rough days seemed to creep back. I also noticed a pattern: on weeks when work was toughest, my dreams were more uplifting, which fits the theory of counterweight dreams." Even if you don't see clear quantitative changes, the qualitative insight can be rich. You might never have connected, say, your snapping in anger (wrath trap) on Friday to the fact you skimped on sleep all week – but the data might reveal that pattern, suggesting the true fix was in ledger maintenance (sleep), not just "anger management" techniques in isolation.

This single-person experiment embodies the spirit of citizen science and the message of this chapter: you are allowed to tinker with your own inputs and observe the outputs. The Law of Fairness gives a guiding hypothesis (that experience tends toward balance by the end), but it's up to you to see how balance attempts manifest in you. By doing so, you become an active participant in understanding your mind. You're not proving the Law true or false with one test on yourself, but you are learning the kinds of evidence it would entail – and perhaps improving your life in the process by adopting a beneficial new habit. In Chapters 22 and 23 (the formal playbook and the hardest objections to LoF), we'll scale this kind of thinking up to groups and more rigorous tests, but it always boils down to honest observation and a willingness to adjust. You've just given yourself a taste of that, on the smallest yet most personal scale.

21.6.5 Where we go next:

We're almost ready to leave the Ledger Gym. But before we do, there's an essential topic to cover. In 21.7, we lay out the *ethical and practical warnings* about what not to do with this knowledge. It's crucial to establish boundaries so that no one misinterprets the Law of Fairness or the Ledger Gym analogy in harmful ways. Let's conclude the chapter by making those "don'ts" crystal clear.

21.7 What Not to Do: Ethical and Practical Warnings

As we close this chapter, it's critical to set boundaries. The Law of Fairness, if true, is a natural constraint – not a tool to be gamed, not a moral loophole, and certainly not a license to harm yourself or others. The Ledger Gym metaphor should never be twisted into self-torture or rationalized cruelty. We therefore highlight a few "Don't" principles to ensure you stay on the right side of both ethics and common sense:

21.7.1 Do not induce suffering to “earn” happiness later

LoF is not a point system where you can deliberately rack up pain expecting an entitlement of pleasure as payback. A reader might wonder, “If life balances out, maybe I can take a shortcut: endure a horrible year now and then coast with good times for the rest of my life.” That’s a dangerous misinterpretation. First, there’s no guarantee how or when compensation arrives – LoF isn’t a vending machine where you insert suffering and joy pops out on demand. Second, you may incur irreparable damage in the process. For instance, purposefully depriving yourself of sleep for weeks could lead to a mental health crisis; inflicting pain or severe isolation on yourself can cause trauma that doesn’t simply wash away with a later high. The ethical stance of this book (echoing Chapter 19) is that suffering is to be *relieved*, not stockpiled or ignored. The Law of Fairness is not a get-happiness-quick scheme; it’s a hypothesis about a boundary condition at life’s end, not a strategy to hack your daily existence. The better approach is to handle the suffering that naturally comes your way with resilience and seek meaning or growth through it – not to seek suffering itself. Think of LoF as a safety net, not as a trampoline to exploit at risk of breaking your neck. In practical terms: Don’t do things that harm you assuming you’ll “get it back” later. Life doesn’t work like banking hours of misery to withdraw joy with interest. Keep living as if your choices and their immediate consequences *still matter* – because they do.

21.7.2 Do not harm others or neglect compassion, ever

Perhaps even more importantly, LoF never justifies cruelty or indifference toward others. One might cynically think, “If everyone’s pain gets balanced eventually, then causing someone pain is no big deal – they’ll get compensated down the line.” This is absolutely wrong. As stated throughout, balance is not a moral excuse. If anything, if you harm someone, *you* may become the instrument of their eventual compensation by needing to atone, or life will compensate them in ways that exclude you (for example, you might lose them as a friend and they’ll find comfort elsewhere while you’re left with regret). In all cases, by hurting someone you’ve only made the journey harder – for them and likely for yourself, since your ledger now carries the weight of having inflicted harm. The ethical guardrail here is clear: “*Relief is a systems variable*.” This phrase from Chapter 19 means

we treat helping and healing as integral parts of the system, not as interference with some cosmic plan. Always offer comfort, always alleviate pain when you can – doing so doesn’t “mess up” any cosmic accounting; in fact, it is the very mechanism by which fairness (if it exists) is achieved. So never withhold kindness under a mistaken belief that suffering is somehow “necessary” to fulfill the law. Your moral compass and basic empathy remain paramount. If someone is drowning, you throw the lifeline now – even if you think things will “balance out in the end,” that’s irrelevant to the immediate moment. In short, do what is right and humane in each situation, LoF or no LoF. The theory lives or dies by natural evidence, not by us forcing scenarios to fit it. We are here to ease each other’s burdens, *not* to create pain thinking we’re helping the universe balance something.

21.7.3 Avoid teleological traps and confirmation bias

When experimenting with your own habits or interpreting life events, be careful not to see *every single thing* as “the universe balancing me.” Humans are storytelling creatures, and there’s a risk of over-fitting the theory to random events. Not every nice thing that happens after a bad day is proof of LoF; not every misfortune after a lucky streak is cosmic rebalancing – sometimes, life just has statistical fluctuations and coincidences. We urged scientific rigor earlier, and that includes recognizing randomness and not reading intentional patterns where there are none. LoF is proposed as a constraint of nature, not an intelligent entity with a plan for teaching lessons. So avoid the mindset of “What is the ledger trying to teach me here?” as if it had a will or moral narrative – that drifts into teleology (the idea that things happen for a purposeful reason). Instead, frame it as “Is there a compensatory pattern I can observe or encourage here?” That keeps you grounded in mechanism and evidence. Also, guard against the *just-world fallacy*: do not assume if something bad happened to someone (including yourself) that it must be “deserved” or caused by some prior imbalance because of LoF. People fall into that reasoning trap easily (“They must have done something to warrant this hardship”). LoF does not operate on a per-event moral desert; it’s a whole-life neutral-sum hypothesis, and even that is unproven. Many completely innocent people suffer immensely – LoF would only hold that somewhere in their lifespan (or mind’s coping processes), counterbalancing positives emerge, **not** that the suffering was warranted or okay. Always emphasize opportunities for repair and support (in yourself and others) rather than assigning cosmic blame. In practical terms: If you catch yourself constructing a neat narrative that *everything* in life is LoF balancing, pause and apply healthy skepticism. Nature is under no obligation to give meaning to each moment – that’s something we humans tend to do. Keep some scientific detachment: look for patterns across time and scale, not in every trivial up or down.

21.7.4 Don't abandon professional help or life goals in favor of “letting LoF fix it”

This is a personal caution for anyone struggling. If you’re dealing with a serious issue – be it depression, anxiety, addiction, a toxic situation, or any life-threatening circumstance – *LoF is not a substitute* for professional intervention or purposeful action. The Law of Fairness, even if true in some ultimate sense, might eventually neutralize your ledger on average, but that’s cold comfort if you’re stuck in a bad place *now*. Use the Ledger Gym ideas as adjuncts to proper care, not replacements. For example, if you’re clinically depressed, developing a morning exercise habit and tracking your mood is great and could help, but you should likely also seek therapy or medical advice. Don’t think, “I won’t take antidepressants or address this trauma because my brain will sort it out naturally eventually.” Maybe it will, maybe it won’t – LoF doesn’t guarantee the timing or magnitude of compensation, and even if it did, *suffering need not be endured passively*. By all means, use the optimism of LoF (the idea that things can change) as motivation to persevere, but still take charge of your well-being through all available means. Similarly, don’t forgo pursuing happiness or success because you fear LoF will “knock you down” if you get too high. Chapter 20 dismantled the false idea that you should hold back from joy to avoid future pain – living fully is not only okay, it’s the point of being alive. The guardrails (if LoF holds) are there to support your journey, not to deter you from it. In short, continue to set goals, seek help, and live with intention. LoF is not fate; it’s a hypothesis that life *tends* toward emotional neutrality, but it doesn’t remove our responsibility to actively improve ourselves and our world.

21.7.5 Keep ethical invariance

This is a fancy way to say: *act with integrity, no matter what*. One of LoF’s tests in Chapter 8 was about measurement invariance – making sure metrics hold across conditions. Here we use “invariance” loosely to mean your ethical standards should remain unchanged whether you believe LoF strongly, weakly, or not at all. If a compassionate act was good pre-LoF, it’s still good post-LoF. If causing harm was bad before, it’s still bad after entertaining this theory. Our theory predicts certain behaviors (like increased reconciliation as one’s horizon shrinks near life’s end), but it doesn’t invent new morals. Interestingly, LoF aligns well with many moral intuitions (for example, that forgiving and making amends is beneficial, or that wanton cruelty boomerangs back), but that’s convergence, not causation. We must be vigilant not to let the concept of a “balancing universe” erode our commitment to doing the right thing now. LoF is never an alibi to be careless or cruel. We state this repeatedly because history shows how easily people can twist any notion of cosmic law into fatalism (“why bother trying if it all balances out”) or into justification for wrongdoing (“I’ll do X because destiny/fate/God

will sort it out"). This book stands firmly against that. The only acceptable uses of LoF in guiding actions are to encourage empathy and patience – empathy because everyone's highs and lows are transient and we're all in the same oscillating boat, patience because if you or someone you love is down, better times likely come with effort and care. But these are just reinforcements of age-old virtues, not new rules.

21.7.6 The Ledger Gym is an analogy, not an exact science of happiness

Finally, remember that the Ledger Gym is a metaphor to help you approach life proactively, not a precise formula for happiness. There will be days that feel utterly *unfair* – days when despite all your good habits and goodwill, tragedy strikes or you just feel terrible for no obvious reason. LoF is not there to blame you for those days, nor to mystically swoop in and fix them on the spot. It is there as a lens suggesting that over a longer arc, you won't be left indefinitely in darkness. But how that arc unfolds is, in part, up to you and your community. By engaging in habits that promote balance, seeking repairs when things break, and avoiding the traps that throw us into vicious cycles, you are doing your part in respecting the hypothetical law – or at least in living a healthier, more stable life. The rest – proving the law, refining its details – we will tackle with proper studies and open debate in later chapters. But the personal insights and precautions from this chapter should serve as your compass in daily life. Science aside, treating life as a ledger to lovingly maintain – with regular audits, clean-ups, and a few “spotters” (friends) to help when the weight gets heavy – is an inherently sound approach to well-being. It won't guarantee a perfect equilibrium every day (nothing can), but it will ensure that when you swing, you swing within safe limits. And when the time finally comes that your life's ledger closes, you can feel at peace knowing you participated fully in the balancing act, with courage and compassion.

21.7.7 Where we go next:

Having built our personal “gym” and ensured we use it wisely, we move forward to more formal terrain. In Chapter 22, we leave the individual scale and the metaphor behind, and step into the real-world lab. We begin in 22.1 by locking in a decision-grade preregistration – that is, crafting a meticulous experimental plan that names the LoF constructs, success/fail criteria, and rules for analysis. This is the first step in the Scientific Playbook that will test LoF with the rigor and transparency it demands. Get ready to take everything we've explored anecdotally and put it under the microscope of hard data and professional scrutiny.

Chapter 22 — The Scientific Playbook

The conference room is full—clinicians, methodologists, engineers, a privacy officer, two skeptics at the back. On the wall: a proposal to test the Law of Fairness (LoF) across a network of hospitals, classrooms, and home telemetry. Someone says, “If we do this, everything goes in writing before we start.” Another: “One-click reruns from raw to figures, or we don’t ship.” A statistician adds, “Invite a red team to try to break it before we unblind.” The room nods. If LoF is going to be tested now, it must be tested at the standard you’d trust with a patient, a publication, or a policy.

This chapter is that standard. The Scientific Playbook turns a brave idea into a research program that can survive adversarial review and support real-world decisions. This chapter will show how to run honest, low-stakes demonstrations—pilot protocols for students, teachers, and citizen-scientists. In this chapter we will scale to decision-grade evidence: larger samples and longer windows; multi-site and federated designs; preregistration that actually binds; analysis plans that remain stable under pressure; and data pipelines so transparent that an independent team—or a regulator—can audit every step.

At its core, this chapter is a public contract. It commits you to work that is easy to rerun, easy to critique, and easy to improve. It also commits you to care: Relief is a systems variable; comfort and dignity override data collection. Nothing in this playbook permits trading someone’s present welfare for prettier data later. That ethic is non-negotiable.

Why this chapter exists. The Law of Fairness is a strong claim about the structure of lived experience. Strong claims deserve strong tests. That means hypotheses that name the constructs (horizon H , channels C , admissible set $\mathcal{A}(t; \bar{L}, H, C)$, shadow price $\lambda(H, \bar{L}, C)$, feasibility Φ), preregistered success/failure gates, and rival models ready to compete on out-of-sample ground. It means analysis plans written so clearly that a capable adversary could execute them and reach the same conclusion. It means measurement that travels across devices, languages, and sites—or honest limits when it cannot. And it means end-of-life work, if attempted, is conducted only with C open and logged and judged against preregistered equivalence margins rather than wishful thinking.

What this chapter demands. We organize the work around five non-negotiables—the Five R’s:

- Registered. Every confirmatory analysis is preregistered, time-stamped, and tied to a specific code commit and container image. Success/failure thresholds are defined in advance; deviations are logged and moved to exploratory analyses by rule.

- Reproducible. From raw to figures is one click in a pinned environment. Containers, checksums, and “golden tests” ensure the same inputs produce the same outputs. No raw PII enters version control; synthetic datasets accompany public artifacts.
- Robust. Sensitivity (“multiverse”) paths are declared up front. You show that signs, intervals, and conclusions survive reasonable choices of filters, links, and random-effects structures. Menu-size models begin Poisson; if over-dispersion > 1.2, you switch to Negative Binomial and state the link.
- Reviewed. An independent methodologist (or team) performs adversarial review before unblinding: code reads, blinded data checks, failure-mode rehearsals. Red-team challenges are encouraged, with clear victory conditions and publication of accepted attacks.
- Responsible. Participants’ rights, safety, and dignity come first. Nulls and negatives are reported. Privacy is engineered (tiered access, k-anonymity/linkage checks, differential privacy where feasible). Community summaries are published in plain speech.

A few additional gates appear throughout the chapter and will be familiar by now. Measurement invariance is required for cross-group claims (configural \rightarrow metric; scalar where powered). If metric invariance fails, you restrict to within-person claims. Rival models are first-class citizens; predictive-coding/free-energy, RL+homeostasis, and adaptation/opponent processes are compared on OOS metrics—WAIC / LOO / log-loss—named once and used consistently. Where end-of-life neutrality is tested, you use equivalence, not difference: windowed mean within ± 0.15 z, slope within ± 0.05 z/day, and variance ratio ≤ 0.80 against a matched baseline—and only when channels (C) are open and logged.

What you will build here. First, an audit-proof preregistration that names the LoF constructs your design manipulates and measures, along with the falsifiers that would make you call the test a bust. Second, a reproducible HCI stack with privacy by design and public Tier-0 artifacts others can run. Third, multi-site/federated replication plans that show your effect is not a lab story. Fourth, adversarial collaboration and red-team bounties that make critique part of the method rather than a post-hoc chore. Fifth, a technical note on end-of-life equivalence (neutral mean, bounded drift, variance compression) so that your most consequential claims are judged against preregistered, humane standards.

What you'll get from this Chapter:

- A comprehensive blueprint for testing the Law of Fairness as rigorous science. The chapter sets the tone with a conference-room scene: experts agree that any LoF study must meet the highest standards. Everything is done by the book – hypotheses and analysis plans are written down in advance, data processing is fully automated and reproducible, and even invited “red teams” probe the design before unblinding. This means you’ll learn exactly how to preregister every decision (locking in success criteria and analysis rules) and how to document every step so that an adversary could reproduce your results with one click. Crucially, these procedures are framed as a public contract: they make your work transparent and critiquable, and they embed an unwavering commitment to participants’ well-being.
- You’ll see how to build a complete preregistration package (Section 22.1), naming every LoF construct in your design and setting hard gates for success. You’ll learn how to set up the Hedonic Composite Index pipeline and open-data hygiene (Section 22.2), including version-controlled code, containers, and privacy safeguards. The text also shows how to plan multi-site studies (Section 22.3), coordinating parallel experiments in different locations so that findings hold up across samples. Adversarial collaboration is stressed, too: Section 22.4 encourages inviting external skeptics to stress-test the experiment and even offering bounties for effective critiques. By the end, you will have a clear understanding of how all these components fit together into a transparent, audit-ready research infrastructure for LoF.
- Equips you to make any test of LoF robust and credible. You’ll internalize the Five R’s framework and know how to apply it: preregistering every analysis, containerizing your code, planning for replicability, and respecting participants at every step. By following the Playbook’s guidelines and examples, anyone applying the Law of Fairness in research will “do it right” – with pre-specified hypotheses, reproducible pipelines, adversarial review, and ethical safeguards as standard practice. In doing so, the chapter makes clear that studying LoF is not just a theoretical exercise but a careful, collaborative science project designed to be trusted by the scientific community.

Subsections in this Chapter:

- **22.1 Full Prereg Packages (drop-in, decision-grade)** - Introduces the comprehensive preregistration kit that will underpin the entire project. This subsection walks through building a complete, decision-grade preregistration

package – locking in hypotheses, analysis plans, and success criteria before any data is collected – to ensure the study’s outcomes will be credible and testable.

- **22.2 HCI Code and Open Data Hygiene (drop-in, LoF-native)** - Lays out how to set up the Hedonic Composite Index pipeline and data practices for maximum rigor and transparency. It covers establishing reproducible code (with version control and containerization), ensuring privacy and open-data standards, and aligning every technical element (data schemas, analysis code, etc.) with Law-of-Fairness requirements so that all collected data can be trusted and shared responsibly.
- **22.3 Multi-Site Replication** - Describes how to design the study for replication across multiple sites from the outset. This subsection details coordinating parallel studies (e.g. in multiple hospitals or labs), including preregistering a shared protocol, synchronizing data collection methods, and planning pooled analyses. The goal is to achieve robust, reproducible evidence by verifying that results hold across different samples and locations.
- **22.4 Red-Team Challenges and Bounties** - Focuses on stress-testing the research via adversarial collaboration. Here we establish methods for inviting “red team” skeptics to find flaws in the design or analysis – even offering bounties or predefined challenges to rigorously probe the study’s assumptions. By building in adversarial reviews and challenge trials, this subsection ensures the LoF experiment can withstand the harshest scrutiny and that any weakness is revealed and addressed early.
- **22.5 Research notes: Equivalence Testing for $L(T)$** - Provides a technical Research note on how to statistically confirm the Law’s predicted balance at life’s end. It introduces equivalence testing methods for the terminal ledger $L(T)$, explaining how one can formally test that the final balance is effectively neutral (within a tight margin) rather than simply failing to reject a difference. This subsection offers the advanced statistical toolkit needed to declare support for LoF’s “neutral ledger” with confidence or to recognize when results fall outside the equivalence bounds.
- **22.6 Classroom Dream Counterweights** - The chapter opens with a classroom-friendly experiment using dreams as emotional “counterweights.” Students will keep a daily journal of their mood and compare it with the tone of that night’s dreams. The idea is to test a key prediction of the Law of Fairness: after a really tough day, do we tend to get especially positive dreams (and vice versa)? This section walks you through a two-week protocol of tracking mood and dream valence, including how to set up a preregistered analysis to catch the expected inversion (bad days followed by good dreams). It’s a hands-on way for beginners

to see if life's ledgers visibly balance – using just pen, paper, and a bit of scientific method.

- **22.7 A Simple Horizon Task** - Next, we introduce an experiment simulating a “last chance” scenario. This section shows how to design a simple decision-making game where some participants think they have many opportunities ahead, and others think it’s their *final* opportunity. By comparing the choices made in these long-horizon vs. short-horizon conditions, students can observe whether people shift toward more meaningful, prosocial actions when they feel time is running out. Clear criteria are provided to determine if the Law of Fairness’s prediction holds in this mini-game, and what results would count as evidence for or against balancing behavior as one’s horizon shrinks.
- **22.8 Ethics and Blinds for Teens** - Any student-led study needs to be done ethically, and this subsection highlights exactly that. It covers how to obtain proper consent and assent when working with minors (your classmates), and how to incorporate simple blinding techniques to keep the experiment fair. The guidance here ensures young researchers respect privacy and avoid bias – for instance, by anonymizing data or not telling participants the hypothesis upfront. By following these principles (minimal deception, confidentiality, voluntary participation), teens can conduct their LoF experiments responsibly and safely. The emphasis is that good science and good ethics must go hand in hand, even in a classroom.
- **22.9 Research Notes: Consent Templates** - The final part of the chapter provides practical tools – sample consent forms and checklists that can be used in small-scale studies. These templates are ready to be adapted for projects like the ones described above. They cover all the basics: explaining the study’s purpose in plain language, outlining what participants will do, noting any potential risks, and guaranteeing confidentiality and the right to withdraw at any time. Essentially, this section gives new researchers a starting kit for paperwork and ethical compliance, so nothing important is overlooked when launching a mini-experiment on LoF.

Where we go next:

In the pages ahead, we leave nothing to chance. First up is Section 22.1, where we construct the preregistration package that will ground your entire project. By nailing that down, you set the tone for everything else in this “Scientific Playbook”.

22.1 Full Prereg Packages (drop-in, decision-grade)

Aim: Make your result predictable before unblinding—and falsifiable. Every element below ties explicitly to LoF constructs: horizon H, channels C, admissible set $\mathcal{A}(t; \bar{L}, H, C)$, shadow price $\lambda(H, \bar{L}, C)$, feasibility Φ , and the life ledger $\bar{L}(t) = \int_0^t \hat{H}Cl(\tau) d\tau$. (LoF is framed as a constraint, not a purpose, so these preregistrations lock in what counts as evidence for or against that constraint.)

22.1.1 Hard gates (pre-declare numeric criteria)

Before collecting confirmatory data, decide what “success” requires. If these gates fail, you’ll treat the result as exploratory. For example:

- Horizon manipulation check: The short-horizon vs. long-horizon condition must produce a perceived time-pressure difference of ≥ 10 points on a 0–100 scale (e.g., a post-task VAS: 0 = “plenty of time”; 100 = “almost out of time”). If this fails, declare the manipulation unsuccessful and report any outcomes descriptively only (not as a test of LoF).
- Utility matching check: In pilot testing, the absolute immediate-appeal gap between paired options must be $|\Delta| \leq 5$ (on a 0–100 appeal rating) for each stimulus pair. This ensures the two choices are equally tempting “in the moment.” If the gap is larger, re-tune and re-pilot the stimuli before launch (and publish those pilot stats).
- Minimally interesting effect size: Decide the smallest effect size that would be operationally meaningful. For example, Δp (the lift in repair-choice probability under short horizon) might be set at +0.04 to +0.05. Any smaller advantage is considered trivial (no practical significance).
- Confirmatory parameter inclusion: If your design can capture both horizon (H) and flexibility (Φ) factors, plan to report the $H \times \Phi$ interaction term in all confirmatory models. (If your study cannot vary both, preregister that limitation.)
- Rival win rule: Declare what it means for an alternative theory to win. For instance: *“If a rival model outperforms the LoF model on all preregistered primary endpoints’ out-of-sample predictive metrics (e.g. lower log-loss or WAIC), then LoF loses this test.”* In other words, you’ll openly admit a defeat if every primary outcome is better predicted by a competitor.

22.1.2 Package contents (auditable)

Every preregistration package should be a complete kit that another team could pick up and follow. Include at minimum:

- Registry capsule: A public preregistration link (or time-locked registry entry) and a version-control commit tag (e.g. commit hash “v1.0-prereg”) tying to the exact code used.
- Protocol document: A PDF describing the design, tasks, how H is manipulated, and how C (channels) are documented.
- Hypotheses (model-addressable): Enumerate H1–H4 directly tied to LoF quantities ($\mathcal{A}, \lambda, \Phi, \tilde{L}(t)$). (See examples below.)
- Outcomes and features: Clearly define primary vs. secondary outcome measures, and pin the analysis code that computes each metric (so there’s no ambiguity in how, say, “ledger variance” is calculated).
- Sampling and power: Justify your sample size via simulation or power analysis (account for clustering with an ICC if relevant) and state the minimally interesting effect (from above).
- Randomization, blinding, unblinding SOP: Describe how random assignment is done (set random seeds, etc.), how conditions are labeled (e.g. masked as “A/B”), the procedure for freezing the data and code, conducting a red-team review (see Section 22.4) while still blinded, and when/how the data will finally be unmasked.
- Analysis plan: Specify all confirmatory models, covariates, and exclusion rules in advance.
- Rivals and comparison: List the competitor models you will pit against LoF (with references or brief description) and reiterate the win rule from above.
- Multiverse grid: If you anticipate researcher degrees of freedom, pre-tabulate a limited set of analyses you’ll check. For example, “up to 3 defensible choices” for each of a few analysis decisions (e.g. which imputation method, which threshold for outlier trials), yielding a grid of analyses you’ll run and report in full. Plan this before seeing data.
- Negative controls and fail-safe tests: List any negative controls (outcomes that should not change if LoF holds, e.g. a “same–same” choice where both options are equally flexible) and any pre-registered Fail patterns. For instance, you might include a checklist of patterns that, if observed, would *falsify* LoF (see 22.5.9).

- HCI measurement plan: Document how the Hedonic Composite Index will be measured or estimated – e.g. calibration routines, a mini invariance check across groups, device/firmware logging.
- Data governance: Outline privacy protections (no direct PII in raw data), how IDs are handled (e.g. salted hashes), and any data retention limits or plans for eventual public release.
- Ethics binder: Include consent/assent forms, a “60-second crisis card” (quick reference for staff if a participant is in distress), and clearly stated red lines (conditions under which the study would be stopped for safety).
- Deviation log: Prepare a CSV or table where any deviations from the prereg plan will be recorded, with timestamps and justifications. If a change is made to the confirmatory plan, it should be logged here and that analysis labeled exploratory in the report.
- Roles and accountability (RACI): List team roles (who is responsible, accountable, consulted, informed for each part). This makes responsibilities clear—e.g. who has authority to halt the study if an issue arises.

As a concrete template, consider a repository structure like this (provided as a drop-in starter kit):

lof-study/

```

|—— prereg/
|   |—— 00_README.md
|   |—— 01_protocol.pdf
|   |—— 02_hypotheses.md
|   |—— 03_outcomes_features.md
|   |—— 04_power_randomization_blinding.md
|   |—— 05_analysis_plan.md
|   |—— 06_rival_models.md
|   |—— 07_multiverse_grid.csv
|   |—— 08_negative_controls.md

```

22.1.3 Confirmatory hypotheses (LoF-native)

Each hypothesis should directly test a LoF “signature.” Only include those your design can truly test (don’t stretch for all four if your data can’t support it):

- H1: Horizon \times Flexibility ($\Phi \times H$). Claim: As the decision horizon shortens, the probability of choosing a repairable or reversible option increases more when flexibility Φ is high. In other words, under time pressure (H low), people favor relief or easily reversed choices, especially if those choices have high Φ .
 - Model: $\text{choice_repair} \sim H + \Phi + H:\Phi + \text{utility_gap}_z + \text{risk}_z + (1 | \text{participant})$, with H coded as SH = 1 and LH = 0 (or an equivalent coding specified in advance).
 - Reject if: $\beta_{(H:\Phi)} \leq 0$, with the entire 95% CI of the interaction term on or below zero across the preregistered multiverse of analyses (i.e., consistently no positive interaction in any defensible analysis).
 - H₂: Repair advantage at matched utility. Claim: When immediate appeal is held constant (two choices tied in “now” utility), a short horizon (SH) leads to a higher probability of choosing the repair-like option than a long horizon (LH) does.
 - Reject if: $\Delta p_{(\text{repair}, SH-LH)} \leq 0$ in the controlled comparison (with CI excluding any positive difference), given the manipulation checks pass (horizon difference and utility match).
 - H3: Menu tightening ($A(t)$ size). Claim: The size of the admissible set $|A(t; L, H, C)|$ shrinks as the horizon H gets shorter (even after accounting for fatigue or task length). People consider fewer distinct paths when time is running out.

- Reject if: There is no decrease (or an increase) in admissible-set size from LH to SH in a valid manipulation. Any finding of equal-or-greater menu breadth under short horizons would falsify this signature.
- H4: “QS-Residual” (unexplained relief signal). Claim: After modeling out known contributors to choice (like utility, risk, arousal), there remains a residual behavior or neural signal aligned with Φ (the “Queue-Systems” hypothesis) – e.g. a preregistered neural correlate in valuation/control ROIs (e.g. vmPFC/ACC), interpreted as correlational, or a latent factor in behavior that correlates with unmet needs. Reject if: The residual is indistinguishable from 0 within preregistered practical bounds. For example, if no significant neural or behavioral variance remains after accounting for standard factors, then there’s no evidence of an extra LoF mechanism at work.

22.1.4 Outcomes and features (pin the math)

Define all measures rigorously:

- Primary outcomes: e.g. binary choice of a repair option (`choice_repair`), the difference in repair choice rates Δp between SH and LH, the size of the admissible set $A(t)$ during a task, a composite “dream reparative themes” score, an end-of-life variance compression metric (defined later in 22.5). Each primary outcome should map to a hypothesis above.
- Feature library (fixed): Define the core features driving your analysis – for instance, *ReliefGain*, *RepairGain*, *HarmRisk* (precomputed valuations of choices), and *OptionFlex* (the Φ proxies for each option). These should be computed in a standardized way (e.g. via a frozen YAML spec or script) for all participants.
- Ledger estimate: Use a formal integration of HCI over time to estimate each person’s cumulative ledger. Denote this as $\hat{L}(t) = \int_0^t \hat{H}CI(\tau) d\tau$.
- In practice, this may come from a state-space model, CFA, or IRT—whatever method is declared. The integration method (Simpson’s rule, trapezoidal, etc.) should be specified and tested. (We use a “hat” to remind that this is an estimate from HCI; the true ledger $L(T) = \int_0^T F(t) dt$, where $F(t)$ denotes instantaneous felt experience, is unobserved.)
- Covariates: Pre-decide any covariates: e.g. participant age, site/lab indicator, trial order, self-reported fatigue or attention ratings, device model/firmware version, recent sleep hours (for dream or EOL analyses), etc. Only include covariates you planned and have theoretical justification for.

22.1.5 Where we go next:

Now that we've solidified a bulletproof preregistration, the next step is to build an equally rigorous foundation for data. In Section 22.2, we turn to the Hedonic Composite Index and data pipeline – ensuring our code, measures, and data-sharing practices are rock-solid, reproducible, and tailored for testing the Law of Fairness at scale.

22.2 HCI Code and Open Data Hygiene (drop-in, LoF-native)

Aim: Make the Hedonic Composite Index pipeline reproducible, privacy-safe, and LoF-interpretable across sites.

Every code artifact should map to LoF constructs: Latent affect signal: $HCI(t) \rightarrow ledger$ $\hat{L}(t) = \int_0^t \hat{H}CI(\tau) d\tau$ (the running total of well-being).

Decision context: Horizon H , Channels C , Admissible set $\mathcal{A}(t; \hat{L}, H, C)$.

Decision parameters: Shadow price $\lambda(H, \hat{L}, C)$ and feasibility Φ (as derived features).

22.2.1 Non-negotiable gates (you must pass these)

To trust the HCI pipeline and data sharing, build in checks:

- Rebuild determinism: Use a single master seed file to govern all randomness (model initializations, train-test splits, bootstraps, synthetic data generation). Any re-run with the same seeds should produce identical outputs (within a tiny tolerance). Specifically, reruns must hash-match “golden” output files; allow at most 10^{-6} differences for scalar results or 10^{-4} RMSE for time-series. If deterministic reproduction fails, fix it *before* proceeding.
- Measurement mini-invariance: Verify that the HCI measurement model holds across groups. At least configural and metric invariance are required across sites and languages (and scalar invariance for key indices if sample size allows). If invariance fails → flag the site/group as non-comparable and restrict claims to within-person effects for that dataset.
- Privacy thresholds: Use explicit preregistered privacy thresholds (e.g., k -anonymity ≥ 10 and maximum estimated re-identification risk ≤ 0.09) for any released data. If you plan to release any detailed individual-level summaries (Tier-1 aggregates), add differential privacy noise with a preregistered budget (e.g., $\epsilon \leq 3, \delta \leq 10^{-5}$).
- Channel bookkeeping: For each participant (or each day, if longitudinal), log whether each channel was open: $C_{\text{analgesia}}$, C_{sleep} , C_{contact} , $C_{\text{translation}}$, $C_{\text{transport}}$, C_{cash} (each as 0/1, or NA if not applicable). If a channel isn’t logged, record why (e.g., “no translation needed in this study”). Any claim of observing LoF signatures must report the channel availability rate in the results. (If channels were mostly closed, a null finding isn’t very informative.)
- Stimulus and feature freeze: Version-control the stimulus bank (e.g. questionnaire or task stimuli) and the code or spec that generates features

(Relief/Repair/Harm/Flex values). Once you start the study, these are frozen. Any change requires bumping the semantic version (and ideally triggers a re-run of pilots). Cross-version comparisons are invalid unless you provide a migration script and demonstrate it doesn't alter results beyond tolerance. In short, no sneaky mid-study tweaks to stimuli or feature calculations.

22.2.2 Repository and schema (with versioning)

Your HCI codebase should be structured and under version control. For example:

```
hci-stack/
  ├── env/      # containers, dependency pins, checksums
  |   └── Dockerfile
  ├── conda-lock.yml
  ├── manifest.lock
  ├── seeds.toml    # single source of RNG seeds
  ├── schemas/
  |   └── v1/
  |       ├── lofdict.yaml  # variable names, units, types
  |       └── migrations.sql # schema up/down migrations
  ├── src/hci/
  |   ├── features.py    # computes Relief/Repair/Harm/Flex ( $\Phi$  proxies)
  |   ├── ledger.py     # state-space model: HCI -> L(t)
  |   ├── invariance.py # CFA/IRT checks for measurement invariance
  |   ├── qc.py         # data ingest QC and exclusion flags
  |   ├── channels.py   # channel extraction and auditing
  |   └── viz.py
  ├── pipelines/
  |   ├── 00_ingest.py
  |   └── 01_deid.py
```

Some key practices: the schema version (e.g. v1) governs the data format and lives under /schemas vX. If you need to change data fields, write migration scripts and increment the version only after all tests pass. Never store raw data in the repo – e.g. have data/raw/ in .gitignore – to avoid accidental leakage of participant data.

22.2.3 HCl math (fully pinned)

Because HCl is an inferred latent variable, preregister all modeling choices:

- Inputs (modalities): Specify which data streams feed into HCI. Examples: momentary self-reports (EMA), physiology (HRV, EDA), behavior logs (activity, mobility), neural signals (EEG/fMRI, optional).
- Preprocessing (declared): Fix any preprocessing steps *a priori*: e.g. “Use 5-minute windows for HRV, 24-hour blocks for EMA; z-score within site; treat missingness as MAR and apply a Kalman smoother or multiple imputation”. These should be in the docs (perhaps in `hci_math_notes.pdf`).
- Observation model: Define the link function from each observed modality to the latent HCI. For instance, you might assume each data source provides a noisy estimate of HCI – e.g. a hierarchical model or a dynamic factor analysis. Declare priors for these links (again, in the math notes or prereg).
- State model: Decide how HCI evolves. Options include a local-level (random walk) model or an AR(1) process for the latent state. Preregister which you’ll use (and the initial seeding, e.g. draw seeds from `seeds.toml`).
- Outputs: The Outcome of this pipeline will be an HCI time series for each participant (e.g., trial-wise or minute-wise $\text{HCI}(t)$ values), plus aggregated measures (daily means, variances with confidence intervals) and finally the numeric ledger $\bar{L}(t)$ per person. These should come with uncertainty estimates (e.g., credible intervals if Bayesian, or SEs via bootstrapping). All of this must be specified and tested (see `test_ledger.py`, etc.).

22.2.4 Where we go next:

With a robust HCI and data pipeline in place, our focus shifts to scale and reliability. Section 22.3 will take the project multi-site, showing how to replicate the study across different locations. The goal is to ensure that the Law of Fairness signals we’re testing aren’t just a one-off – they should appear consistently, no matter where we look.

22.3 Multi-Site Replication

Aim: Determine whether LoF signatures are portable across sites, languages, devices, and cultures—and if not, pinpoint where they break. Every element here ties to LoF constructs:

- Horizon H
- Channels C
- Admissible set $\mathcal{A}(t; \bar{L}, H, C)$ (we include \bar{L} to acknowledge ledger differences)
- Shadow price $\lambda(H, \bar{L}, C)$
- Feasibility-of-compensation Φ (the same Φ , defined as before)
- Ledger estimate $\hat{L}(t) = \int_0^t \hat{HCl}(\tau) d\tau$ (treating HCl as a measured proxy; “hats” denote estimates).
- (Recall, $L(T) = \int_0^T F(t) dt$ is the true underlying ledger, not directly observed.)

22.3.1 Estimands (declare up front)

What effects will you actually estimate? Preregister the primary and secondary endpoints:

- Primary outcome (behavioral): The pooled *within-person* $H \times \Phi$ interaction effect on repair choices at matched immediate utility. Typically this is tested with a mixed-effects logistic model (random effects per participant and per site). In plain language: do people choose more “repair” options under short horizon than long horizon, especially for high- Φ options, consistently across sites?
- Secondary outcomes: You might include:
 - (1) Menu tightening: the change in admissible set size $|\mathcal{A}(t; \bar{L}, H, C)|$ as H shrinks.
 - (2) HCl responsiveness: how the HCl responds to controlled perturbations (e.g., does +60 min extra sleep produce a predicted HCl boost the next day?).
 - (3) Dream counterweights: the within-person link between daytime strain and reparative themes in that night’s dreams (are “hard days” followed by compensatory dream content?). Define each with specific metrics.
- Heterogeneity: Plan to examine between-site variance and other moderators. For example, estimate between-site variance τ^2 (and I^2 heterogeneity), and test moderators like language, device type, age band, or overall channel-open coverage. A high τ^2 or I^2 might indicate the effect differs by context.

22.3.2 Site pass/fail gates (each site must hit these)

Every site in a multi-site study should meet basic quality thresholds:

- Horizon manipulation strength: At that site, the mean perceived time-pressure difference (SH minus LH) should be ≥ 10 on the 0–100 scale (measured immediately after tasks). If a site’s local “short vs. long” manipulation is weak (e.g. participants didn’t actually feel time-pressured), that site cannot confirm the primary effect.
- Utility matching check: Each site should pilot-test its stimuli to ensure $|\Delta\text{appeal}| \leq 5$ (0–100) for the choice pairs. If one language version or culture finds one option much more appealing than the other on average, you need to adjust and re-pilot at that site before collecting main data.
- HCl mini-invariance: Test measurement invariance for the HCl model at the *measurement level* (not just comparing raw HCl values). You need configural and metric invariance across sites/languages (scalar if you have power for it). If this fails for a site → restrict that site’s data to within-person comparisons only and flag it as non-comparable in aggregated analysis. (You might still include it for exploratory or meta-analytic purposes but treat cross-site differences with caution.)
- Channel logging: As in single-site, each site must log the six channels for each relevant period (e.g. daily). You will report the fraction of time each site had all channels open. If a site systematically has a certain channel closed (e.g. a location where translation help isn’t available, so “contact” channel often fails), that’s important context.
- Privacy/provenance: Each site must adhere to data privacy thresholds (k -anonymity ≥ 10 , linkage risk ≤ 0.09). Ideally, all sites apply differential privacy for any aggregate exports ($\epsilon \leq 3$, $\delta \leq 10^{-5}$), but if a site cannot for legal reasons, they must document why. Also, each site should implement triple-hashing of provenance info: record the exact code commit hash, container digest, and data snapshot ID used for its results.
- Consequence: Sites that fail any gate above can still contribute to exploratory analyses or methods papers, but they are excluded from confirmatory pooling. (Preregister how you’ll handle any failed sites.)

22.3.3 Network architecture (preregister one approach)

Decide how data will be combined *before* you begin, and stick to it:

- Federated model (preferred for sensitive data): Each site runs the analysis in a container locally and only shares summary statistics or model gradients. No raw sensitive data leaves the site.
- Secure pooled model: Tier-1 de-identified data (basic anonymized outcomes) are uploaded to a central secure data safe; highly sensitive data (neural images, geo-location, etc.) remain at sites. The central analysis then works on the pooled anonymized dataset.

Either way, enforce a one-click “dry run” of the entire pipeline (with synthetic data) before the first real participant is run. Capture all provenance hashes at this stage too. The goal is that when real data comes in, each site can essentially press “Go” and the same pipeline runs that was tested with fake data.

22.3.4 Harmonization before first participant

For a multi-site study, spend time upfront to harmonize protocols:

- Stimuli: Freeze the stimulus set version and ensure identical timing, instructions, and inter-trial intervals across sites. (Translation differences aside, participants should essentially be doing the same task everywhere.)
- Local utility match: Have each site do a small pilot (~10–15 participants, or as needed per language) to check that stimuli are equivalently appealing. Publish or share those appeal stats so it’s transparent that all sites achieved the $|\Delta| \leq 5$ utility gap before starting.
- Sensors: Create a device whitelist and pin any firmware versions for wearables or medical devices. For example, decide that only specific models of EEG or a specific actigraphy device firmware are allowed. Standardize settings (e.g. HRV window = 5 min, actigraphy epoch = 30 s). Provide a calibration script or procedure and have each site run it, uploading their calibration logs.
- Language: If multiple languages, do proper forward/backward translation of all participant-facing materials, and cognitive debriefing to ensure questions mean the same. Version each language file (e.g. “Survey v1.0 – Spanish”).
- Training: Train all site staff and verify consistency. For any subjective coding (like dream content coding), require a reliability check (e.g. inter-rater $\kappa \geq 0.80$ on a training set) before they begin coding real data.
- Dry run: Do a “dress rehearsal” with synthetic data (e.g. simulate 10 fake participants per site) to test the entire data pipeline and analysis. All checks from

Section 22.2 (invariance, privacy, golden tests, etc.) must pass with this synthetic run before enrolling real participants.

22.3.5 Sampling and power (simulate where variance lives)

Key design parameters for multi-site:

- Design: Typically a three-level design (trials within participants within sites). Account for this in power calcs.
- Sample size targets: For behavioral-only endpoints, aim for at least ~60 participants *per site*. If you include a heavy neural/physiological module that's costly, you might target ~30 per site (relying on within-person contrasts for power).
- Simulation assumptions: Use plausible values: e.g. baseline "repair" choice probability ~0.5; minimally interesting effect $\Delta p = 0.05$ (5 percentage points); ICC (participant-level) ~0.10–0.20; between-site variance τ^2 consider scenarios like {0, 0.02, 0.05}. Simulate under these to see power.
- Diversity: Ensure your network isn't homogenous. For example, require that the sites collectively cover at least two non-overlapping age bands (e.g. both younger and older populations) and at least one cross-language or cross-culture contrast. This isn't just for generalizability but to truly test transportability of LoF. (Preregister these as sample goals.)

22.3.6 Randomization, blinding, preregistration

- Randomization: Counterbalance horizon order (half participants do SH first, half do LH first, or randomize order) and use the pre-set random seeds to generate assignment. Save those seeds to your provenance record.
- Analyst masking: Keep condition labels masked (e.g. call them "cond_A" vs. "cond_B") for as long as possible – ideally until after an independent red-team review of the analysis (Section 22.4). Also freeze the stimulus bank and feature definitions at prereg (semantic versioning as noted).
- Public preregistration: Post the full analysis plan publicly (or in a time-stamped registry) before collecting data. Include H1–H3 hypotheses, the numeric gates (from 22.3.2), exclusion rules, equivalence margins for replication success (see 22.3.8), the rival models and win criteria, and the path to your deviations log. This way, everyone knows the rules of the game in advance.

22.3.7 Confirmatory models (analysis stanzas to run)

Plan the exact statistical models for confirmatory analysis. For example:

- Behavioral outcome (repair choice): A mixed-effects logistic regression, e.g. $\text{choice_repair} \sim H + \Phi + H:\Phi + \text{utility_gap_z} + \text{risk_z} + C_{\text{open_any}} + (1 | \text{participant}) + (1 | \text{site}) + (H | \text{site})$. Here, H is an indicator for short-vs.-long horizon, and $C_{\text{open_any}}$ may be a covariate noting if any channel was closed for that participant (or a per-trial measure if needed). A random slope for H by site is included to absorb site-level differences in the manipulation effect. (If your preregistration decided to simplify and use intercepts-only, state that explicitly.)
 - Menu tightening (count model): A regression on the size of the admissible set, e.g. $A_{\text{size}} \sim H + L + \text{fatigue} + (1 | \text{participant}) + (1 | \text{site})$. Use a Poisson regression initially. Check for over-dispersion; if the dispersion > 1.2, switch to a negative binomial model and report that you did so. (The variable $\mathcal{A}(t; L, H, C)$ must be clearly defined per task—for example, count of distinct activities the person is pursuing in a given interval.)
- HCl perturbation (e.g. +60 min sleep): If you included a planned intervention to test responsiveness of HCl, model the change: e.g. $\text{delta_HCl} \sim \text{perturbation} + (1|\text{participant}) + (1|\text{site})$. This tests if adding 60 minutes of sleep (or a reconciliation phone call, etc.) shifts the next-day HCl upward, on average.
- Dreams (day-to-night link): e.g. $\text{reparative_composite} \sim \text{day_strain_z} + \text{sleep_hours_z} + (1|\text{participant}) + (1|\text{site})$. This tests if the day's strain predicts the presence of reparative themes in that night's dreams (with perhaps sleep hours as a covariate). You'd expect a positive association if LoF holds. Also include a negative control here: for instance, $\text{threat_theme} \sim \text{day_strain_z}$ should show no positive relationship (stress shouldn't increase negative/threat dream content if the theory is correct and only reparative content increases). Negative controls should remain flat (null) if LoF's mechanism is specific.
- Rivals (for adversarial comparison): Specify alternative models you will run in parallel. For example:
 - *Predictive-coding/free-energy model* – perhaps a model where affect is driven by prediction errors weighted by uncertainty (instead of a fairness ledger).
 - *Reinforcement learning + homeostasis model* – e.g. outcomes driven by maintaining a set-point.

- *Adaptation/opponent-process model* – where initial shocks are opposed by delayed reactions. Each rival should be implemented and *preregistered*. All these models are fit to the data as well.
- Win rule (from preregistration): If a rival model outperforms LoF on the predefined OOS metrics for *all* primary endpoints, we declare that rival the winner for this dataset. For example, if the predictive-coding model has consistently lower log-loss (or WAIC) than the LoF model on every primary outcome, then LoF did not hold up in this replication. State that outcome plainly. (*Note: matched model complexity and regularization across models is important for a fair comparison.*)
- Interference checks: Because data are clustered in time and place, add robustness checks for interference. Use cluster-robust standard errors, include terms like site×time interactions (to capture any network-wide drift), and even do a leave-one-site-out meta-analysis to see if any single site is driving the effect.

22.3.8 Equivalence and non-inferiority (replication is two-sided)

Preregister equivalence margins for key effects – replication isn’t just about $p < .05$ in the same direction; we want to confirm effects are not *meaningfully* smaller than expected either. For LoF, set default margins equal to the minimal interesting effects (and measurement noise) from earlier:

- $H \times \Phi$ interaction (log-odds scale): ± 0.10 on the interaction coefficient. (This corresponds to roughly $\pm 2.5\text{--}3$ percentage points on the probability scale if baseline $p \approx .5$).
- HCl perturbation effect: ± 0.15 SD on the HCl change (for a +60 min sleep or similar intervention).
- Menu tightening: ± 0.05 in the proportion of activities (or options) in $|\mathcal{A}(t; \bar{L}, H, C)|$. Declare that the replication is successful if, for example, the 90% CI for the pooled effect lies entirely within the above equivalence margin (and heterogeneity $I^2 \leq 40\%$). Alternatively, if the effect is in the predicted direction and the point estimate exceeds your minimally interesting threshold (with CI not crossing zero), that can count, so long as it’s not weaker than the margin. Negative controls must still show no effect. In other words, either we see a clearly positive result or a statistically equivalent result—anything in between is inconclusive.

22.3.9 Missingness, exclusions, robustness

- Exclusions (preregistered): Define criteria to drop data: e.g. participants who fail the horizon manipulation check, those with < 60% usable trials (due to attention

fails, etc.), any device failures or safety-related early stops. These will be removed from confirmatory analysis and noted.

- Missing data: Plan how to handle missing responses. For questionnaires, multiple imputation might be used; for continuous HCI time-series gaps, use state-space smoothing that propagates uncertainty (so missing stretches widen the CI). Do *not* cherry-pick methods post hoc—decide now.
- Robustness set: List which robustness analyses you'll run: e.g. leave-one-site-out re-analysis (to ensure no single site drives the effect), analyze only a subset of device types, analyze only one language subgroup, add a dummy variable for study order, etc. These checks probe if the result holds under slight perturbations. (They're not all confirmatory, but you plan them ahead to avoid bias.)

22.3.10 QA/QC during data collection (automated alerts)

Especially in multi-site projects, set up a dashboard or script for ongoing quality checks:

- Daily checks: Monitor completion rates, reaction time distributions (flag if they drift beyond ± 2 SD of the rolling mean), sensor uptime, etc. Trigger an alert if something goes off (e.g. one site's reaction times suddenly get much faster—could indicate a protocol issue).
- Weekly checks: Ensure each site uploads hardware calibration logs or other weekly maintenance. If a site misses a calibration, automatically pause new data collection there until resolved.
- Monthly checks: For any subjective ratings or coding tasks, do rater drift checks. Calculate inter-rater reliability (target $\kappa \geq 0.75$ each month); retrain or recalibrate if it drops below that. Also review the incident log (see below) monthly.

22.3.11 Governance, privacy, ethics

Local IRBs + umbrella oversight: Each site should have its own ethics or IRB approval, and there should be an umbrella data use agreement that all PIs sign onto (covering data sharing rules, publication rights, etc.).

- Data tiering and privacy: Apply the same tiered-data approach as Section 22.2 (Tier-0 public aggregate vs. Tier-1 protected individual data). Enforce differential privacy and k-anonymity as applicable across the network. No free-text data leaves a site (e.g. dream journal text should be coded at the site, not shared raw).
- Incident response: Agree on a 24-hour window for reporting any adverse events or major data incidents to the whole network. If something goes wrong (participant

distress, data breach, etc.), pause data collection at all sites, investigate the cause, and circulate a “root cause analysis” memo with what will be done. Also, ensure no PII ever enters research CSVs.—if an identifier slipped in, that’s a data incident.

22.3.12 Reporting skeleton (uniform and audit-ready)

Plan the structure of your final report so nothing is forgotten. For example:

- Construct audit table: A summary table of manipulation checks and key process metrics: e.g. mean time-pressure ratings for SH vs. LH at each site, $|\Delta\text{appeal}|$ pilot values for each stimulus pair, HCl invariance test results, and channel-open rates by site. This shows that all prerequisites were met (or notes where they were not).
- Methods appendix (CONSORT-R): A CONSORT-style appendix listing each site, language, and device class; stimulus version numbers; container digests and commit hashes run; schema version used, etc. Essentially a reproducibility blueprint.
- Flow diagram: A diagram of participant flow at each site (how many screened, how many enrolled, how many completed, how many excluded from analysis, etc.), analogous to CONSORT flow charts.
- Tables: Prepare tables for confirmatory results (with pooled estimates and site-level random effects if applicable), a table comparing rival models on OOS metrics, and tables for any control conditions (e.g. showing that “same–same” trials produce no difference). Also include a measurement table showing invariance statistics.
- Heterogeneity: Include a section or table for between-site heterogeneity: report τ^2 , I^2 , and any moderator analyses that were pre-specified.
- Fail-pattern registry: Include the checklist of possible Fail patterns (from your prereg plan) and tick which ones (if any) occurred. For instance, mark if “no $H \times \Phi$ effect with successful manipulation” happened, or “EOL compression absent even with channels open,” etc.
- Deviations log: Append the deviations CSV or a table summarizing any deviations from the prereg plan, with timestamps and reasons.
- Data/code availability: State where the data and code can be accessed. Tier-0 (fully anonymized summary data) might have a DOI for a public repository; Tier-1 (detailed data) might be available in a secure enclave for qualified researchers.

22.3.13 Success and failure (predeclared)

It's good practice to decide in advance what outcomes will lead you to conclude "LoF replicated" versus "LoF failed here." For example:

- Success (any one of these): (a) The pooled $\Phi \times H$ effect is statistically significant in the predicted direction, and at least 75% of sites show the effect in that direction (with $I^2 \leq 40\%$ heterogeneity); or (b) equivalence is demonstrated—all sites' effects are within the ± 0.10 (log-odds) band and $I^2 \leq 40\%$ meaning no site has a meaningfully smaller effect — and negative controls remain flat, and none of the rivals outperform LoF on the OOS metrics. (Either clear significance or a thorough equivalence counts as success.)
- Failure (actionable): If any of the following occur, we consider the LoF hypothesis to have *failed* in this domain:
 - All manipulation checks passed, but the pooled $\Phi \times H$ effect is ≤ 0 (no positive effect at all).
 - Every preregistered rival model beats the LoF model on all primary endpoints (meaning LoF added no predictive value).
 - A Fail pattern occurred under open channels (e.g. no horizon effect but channels were open, or other disconfirmatory pattern from your registry).
 - Heterogeneity is extreme: $I^2 > 60\%$ and cannot be explained by pre-specified moderators. In any such case, the action is to downgrade or reject the law-level claim for that domain (and publish the result with a conclusion). Still publish all methods and tools, so others can learn.

22.3.14 Where we go next:

Having designed a multi-site study that can replicate results, we're ready to expose our plan to serious scrutiny. In 22.4, we invite the toughest tests — red team challenges and adversarial audits — to actively probe for weaknesses. This next part ensures that if our study is going to fail, it does so fast and transparently, or else it emerges battle-tested.

22.4 Red-Team Challenges and Bounties

Purpose: If the Law of Fairness is real, it should survive skilled attacks; if it isn't, we want those attacks to succeed (quickly and publicly). To that end, we have an open red-team program with clear targets, ethical guardrails, submission rules, bounties for successful falsifications, and a process to update our claims when someone finds a crack.

(*LoF constructs relevant in this section remain the same: horizon H, channels C, admissible set $\mathcal{A}(t; \bar{L}, H, C)$, shadow price $\lambda(H, \bar{L}, C)$, feasibility score Φ , and the ledger $\bar{L}(t) = \int_0^t HCl(\tau) d\tau$. HCl is our measured proxy; recall that $L(T) = \int_0^T F(t) dt$ denotes the true underlying ledger that we cannot observe directly.*)

22.4.1 What “winning” against LoF looks like

A *red-team win* is any preregistered attack that, under blinded analysis and independent audit, demonstrates one of the following outcomes on LoF's confirmatory endpoints (each corresponds to a key LoF signature being broken):

1. No $\Phi \times H$ effect: In a well-powered, within-person experiment with matched utility and a successful horizon manipulation (SH-LH ≥ 10 on the VAS), the confirmatory model shows no positive interaction. For example, the model choice_repair ~ H + $\Phi + H:\Phi + \dots$ yields $\beta_{(H:\Phi)} \leq 0$, and its CI excludes even the minimal expected effect size (see equivalence margins in 22.3.8). In short, the horizon \times flexibility effect is statistically zero or reversed.
2. No menu tightening: The size of the admissible set does *not decrease* as H shortens (or, worse, it increases). After controlling for fatigue and ledger differences, $|\mathcal{A}(t; \bar{L}, H, C)|$ under short horizons is equal to or larger than under long horizons. This means no evidence of the predicted narrowing of options (or an opposite effect).
3. HCl non-transportable: The HCl measurement model fails to generalize. Specifically, your attack shows that the HCl's latent structure is not invariant across a pre-specified group contrast (e.g. different language or device types) *despite* adequate sample size. In addition, when you force the model to be the same across groups (assuming configural invariance holds but metric fails), the within-person effects flip or vanish. In other words, LoF's core metric doesn't travel well.
4. Rival model wins cleanly: You demonstrate that a specific rival theory can explain the data *better than LoF does*. For example, a predictive-coding (free energy) model, or a pure reinforcement-learning + homeostasis model, or an adaptation

opponent-process model, is pre-registered and ends up outperforming the LoF model on all primary outcomes' OOS metrics. Crucially, this must be with matched model complexity/regularization. If your rival beats LoF on (say) log-loss and WAIC for every primary endpoint, with no caveats, that is a clear falsification of LoF's supposed explanatory power.

5. Boundary-condition violation (with open channels): You find a condition at end-of-life or similar extreme scenario where LoF should apply, but it doesn't. For instance, you show that when all six channels are open and documented, the expected end-of-life pattern fails: the variance of the ledger does not compress (EOL variance stays high or even increases), or “dream counterweights” don't appear at all across individuals and sites. This would mean the supposed fail-safe mechanisms (like relief dreams or narrowing focus) aren't actually universal, even under conditions favorable to LoF.
6. Data-hygiene breach that alters conclusions: You find a flaw in the HCI pipeline or replication kit that isn't just a nitpick but *when corrected, overturns a prior result*. For example, perhaps the HCI calculation had a bug that, once fixed, makes a previously significant LoF effect disappear. Or a data inclusion mistake that, when remedied, nullifies the outcome. This is essentially a quality-assurance win: it shows a past LoF finding was an artifact of a preventable error. (Minor improvements or cosmetic changes don't count; it has to change a conclusion.)

22.4.2 Attack surface map (pick a target, not a vibe)

Would-be red-teamers should focus on specific weak points in the LoF framework, rather than vague “try everything” approaches. Some example attack surfaces and ideas:

- A. Measurement and Modeling: Go after the HCI or ledger construction. For example, break the mini-invariance check – show that the HCI measurement fails even basic configurational or metric invariance across groups. Or show that a particular choice of ledger integration or smoothing (e.g. how missing data is handled) is hiding a negative result; when you alter those choices (within reason), the sign of an effect flips. Another angle: demonstrate that the Φ feature isn't truly monotonic (e.g. the “flexibility” scores for options violate their intended ranking under some conditions), undermining interpretations.
- B. Experimental design: Target the assumptions of the horizon experiments. For instance, show that you can get a large perceived horizon difference (people feel rushed) without the predicted menu tilt – perhaps everyone just rushes through both high- Φ and low- Φ options equally, meaning our manipulation raised arousal

but did not trigger compensatory behavior. Or show that our “same–same” negative control (two equally flexible options) isn’t staying flat – if even identical options lead to different choices under time pressure, it suggests a confound like general arousal or speed-accuracy tradeoff is at play, not the LoF mechanism.

- C. Adversarial prediction: Take the rivalry seriously – train the alternative models and see if they can actually predict better. Using the provided seeds and data splits, fit the preregistered rival models and beat the LoF model on the agreed OOS metric. Another approach: construct a scenario where LoF *should* be silent (a “pure motor” task with no meaningful hedonic stakes) and show that our LoF analysis still finds spurious patterns (a false positive), indicating it isn’t properly controlled.
- D. Privacy and provenance: Without touching any real identities, prove that our data releasing scheme has a flaw. For example, mathematically demonstrate that the combination of age, condition, and a coarse location in a Tier-1 dataset could allow re-identification with > 10% chance – thereby violating our ≤ 0.09 linkage risk rule. Or show that our “rebuild determinism” standard fails: maybe our pipeline isn’t as deterministic as we claim (the “golden tests” don’t always pass on different machines), revealing potential hidden bugs.
- E. Interference/coupling: Find an inter-site or temporal coupling we ignored. For example, show that sites might influence each other (perhaps through staff communications or even participants on social media). If you can demonstrate an unmodeled site \times time interaction that either creates an artificial LoF effect or hides a real one, that’s a serious issue. E.g., maybe overall stress was decreasing over time across all sites (some external event resolving), and that coupled with horizon timing could fake a “horizon” effect unless modeled. Demonstrating such a confound undermines naive LoF analyses.

22.4.3 Ethics and safety guardrails (non-negotiable)

Even adversarial tests must follow strict ethical rules:

- No human harm; no targeting individuals. Attacks must not cause any participant physical or psychological harm, and absolutely no attempts to re-identify participants or breach confidentiality are allowed. Privacy attacks should use theoretical proofs, audits, or synthetic data reconstructions – never trying to expose real identities.
- No deception of participants or staff. You can be creative in analysis or design, but you cannot secretly manipulate participants or trick site staff. All participant

interactions must remain honest and within the approved protocol. (Your adversarial moves happen behind the scenes in the data processing or design structure, not as undeclared interventions on people.)

- Respect governance and infrastructure. Don't break into servers, scrape unpublished data, or bypass established data safeties. If data is in a safe enclave, don't try to brute-force it. We welcome clever analyses, not illegal or unethical access.
- Care first. If at any point you sense participant distress or risk (e.g. your site is seeing adverse events), you must halt and inform the appropriate oversight (see 22.8.4 on crisis protocol). Remember: relief of suffering and dignity always override data collection, even in an adversarial test.

A violation of any of the above disqualifies the submission and could trigger reporting to authorities or IRBs. In short, “red-team” does not mean “no rules”—it means a *different role* with the same fundamental ethics.

22.4.4 Bounty tiers (USD rewards, paid upon audit)

We offer escalating bounties for successful challenges, proportional to their impact:

- T0 – Documentation and reproducibility: Find and fix a purely technical flaw that could lead to errors (e.g. a CI configuration that allowed silent data drift, a schema error causing misalignment). *Example:* You provide a continuous integration recipe or schema correction that catches a potential analysis slip, it passes our audit and we merge it. Payout: \$500–\$2,000.
- T1 – Hygiene and privacy: Expose a data hygiene or privacy oversight. *Example:* You demonstrate that under certain quasi-identifier combinations, our public data release exceeds 0.09 re-identification risk (without actually re-identifying anyone), or that our differential privacy accounting was misapplied. Payout: \$2,000–\$5,000.
- T2 – Measurement: Undermine the measurement model. *Example:* You show that the HCI measurement fails configural or metric invariance across two languages or device types we said it would cover (and you have enough sample to be sure). Or you identify a systematic Φ feature mis-ordering. Payout: \$5,000–\$10,000.
- T3 – Design falsifier: Achieve a clear null result on a core behavioral signature despite proper conditions. *Example:* You run a well-powered study where $\Phi \times H$ should appear ($\text{manipulation} \geq 10$, $|\Delta\text{Appeal}| \leq 5$ to ensure a fair test), and you get no effect (or opposite effect) with tight CI. Payout: \$10,000–\$20,000.

- T4 – Rival model win: Demonstrate a rival theory that out-predicts LoF across the board. *Example:* Your predictive-coding model or RL-homeostasis model (with complexity similar to ours) beats the LoF model on all preregistered OOS metrics in the horizon tasks and maybe also in dream analysis. Payout: \$20,000–\$40,000.
- T5 – Boundary-condition fail: Show that when all guardrails are open, LoF’s predicted endgame effects don’t happen. *Example:* In a multi-site end-of-life dataset where analgesia, contact, etc., are assured, you find no ledger neutrality or no dream counterweights—variance doesn’t compress or dreams don’t invert suffering. Payout: \$40,000–\$75,000.
- T6 – Paradigm-level refutation: Present a general refutation of LoF. *Example:* A theoretical argument plus evidence across studies that every LoF signature can be explained by some other combination of known processes (adaptation, coping, etc.) with no need for a “fairness” construct. In other words, LoF reduces to an existing theory and adds zero predictive power. Payout: Negotiated case-by-case (likely very high) + co-authorship on a high-profile paper or correction.

(Bounty ranges may be adjusted based on scope — e.g. a multi-site or adding neural measures tends toward the high end. Some challenges might be pre-funded with fixed rewards.)

22.4.5 Submission workflow (what to send, how we judge)

We require a structured submission for any challenge attempt. Here’s the process:

1. Register your intent: Before you start the attack, submit a brief registration (timestamped) describing your target (which tier/hypothesis you’re testing), the planned dataset or context, and a high-level plan of attack (including which analysis you’ll preregister). We just log this to prevent after-the-fact claims.
2. Freeze your artifacts: Once you’re ready to execute, lock down everything. Record the container digest (if you built new code, containerize it), commit hash of your analysis code, schema version you’re using, and a hash of your seeds file. This ensures you can’t tweak things after peeking at data.
3. Run blinded (if applicable): If your test involves running on LoF data that has conditions, we can provide you with masked labels (cond_A vs. cond_B) or synthetic rehearsal datasets. You perform your analysis without knowing which is which. If your attack is more theoretical (like proving a privacy leak), this step might not apply.

4. Deliver the bundle: Submit your code, a manifest of everything needed to reproduce the results, an analysis notebook or report, the outputs, and a concise 1–2 page “attack memo.” The memo should state: your hypothesis (what specific failure you aimed to show), methods, the pass/fail criteria you set, and which LoF claim is at risk if you succeed.
5. Audit: An independent reviewer (or our team’s designated auditor) will re-run your analysis in a fresh environment. They will check all the prerequisites: e.g. did the horizon VAS actually reach ≥ 10 SH–LH? Were $|\Delta\text{appeal}| \leq 5$? Did you respect the over-dispersion rules and invariance gates? Did all golden tests pass (i.e. your pipeline is deterministic)? And of course, do your results replicate on the auditor’s side, and do OOS metrics support your claim?
6. Decision: Based on the audit, we either accept the challenge (and award the bounty), reject it (with detailed reasons—perhaps a gate was failed or the effect was not as claimed), or request additional analysis if something is unclear. All accepted attacks will be made public (with a DOI), and payouts occur within 30 days of acceptance.

Also note disqualifiers (as mentioned in 22.4.3): any evidence of p-hacking beyond your prereg plan, undisclosed multiple testing, switching endpoints after seeing results, or using unapproved data will invalidate the submission.

22.4.6 LoF-provided red-team kit

We want to make it as straightforward as possible for someone to try to break LoF (in a valid way). To that end, we provide a kit to interested challengers:

- Prebuilt containers that include all our frozen analysis code from Chapters 22.1–22.3, plus baseline implementations of rival models, and the seeds used in our main studies. Essentially, you can pull a Docker image that has everything needed to rerun our analysis and then modify one aspect for your attack.
- Golden tests and synthetic data: We share synthetic datasets (with known ground-truth H and Φ effects injected) so you can test your setup. Golden test cases are included to ensure your environment is producing identical outputs for identical inputs (so you don’t inadvertently claim a difference that was just a glitch).
- Stimulus banks: Access to our standardized stimulus sets (with version history) along with the pretest ratings of their immediate appeal. This helps you design any new experiments without having to start from scratch on stimuli.

- HCI measurement spec: Detailed documentation of the HCI model (priors, link functions, missing data handling) and the invariance scripts we use, along with the criteria for passing. Essentially, we hand over how we build HCI and what counts as “acceptable” invariance.
- Privacy linter: A tool or script that checks a dataset against our privacy rules (calculating k-anonymity, linkage risk, DP usage). You can run this on your results to ensure you’re within bounds before submission. (If a site chooses not to apply DP, they must explicitly justify it; the linter will flag if DP is absent so that justification is required.)
- Construct-audit templates: Blank templates for the tables and checks we expect (manipulation check results, channel coverage, invariance outcomes). You can fill these in with your attack’s values to show how things compared to the original.

22.4.7 What happens if you win (and if we do)

- If the red team wins: We will (a) publish the attack and a reproduction of it in our own pipeline (transparency), (b) downgrade or correct the LoF claim in question (e.g. annotate the margin of the chapter and maintain an online errata page explaining the issue), (c) update our analysis kits, documentation, or methodology to ensure future studies aren’t vulnerable to the same issue, and (d) invite the challengers as co-authors on any formal correction or methods paper that results. In short, a win triggers a public correction and improvement cycle.
- If LoF survives the attack: We will still publish the attempt and a detailed explanation of why it *didn’t* succeed. Perhaps the manipulation was insufficient, or invariance failed in the attacker’s data, or the rival model overfit – whatever the reason, it’ll be documented. The red team gets credit for the attempt, and importantly, if in the process they improved our tools (say, they found a documentation bug – a T0/T1 issue), we will pay out that corresponding bounty anyway. In other words, even an unsuccessful Tier-3+ challenge can earn a smaller reward if it helped us make the system more solid.

22.4.8 Exemplars (challenge shapes to copy)

To spur ideas, here are hypothetical examples of high-quality challenges:

- Equivalence-first replication: Instead of testing “is there an effect,” design a replication that tests *equivalence*. For instance, set a ± 0.10 log-odds ROPE for the $\Phi \times H$ interaction (the core effect) and show that the 90% CI falls entirely within [-

$0.10, +0.10]$, with $I^2 \leq 40\%$ and all negative controls flat. This would count against LoF in a serious way – it suggests the effect, if any, is so tiny it's practically zero.

- Over-dispersion trap: Construct a scenario where using the wrong model gives a false positive. For example, create a dataset where the admissible set size does not actually change, but is highly variable. If one naively uses Poisson, they might see a “significant” result; you then show that the Poisson model was over-dispersed (say dispersion = 1.5) and switching to a negative binomial with the proper link makes the effect disappear. This highlights a methodological fragility in analysis choices.
- Coupling confound: Find or simulate a coupling that we didn't account for. For example, maybe participants who started in the SH condition talk to those who haven't done it yet (site \times time coupling), creating a bleed-over effect. Demonstrate that this happened and that once you model or eliminate this interference, the LoF effect vanishes. This shows the original result was an artifact of not accounting for a certain dependency.
- Rival supremacy: Take one of the rival models and really tune it well (without overfitting). Perhaps you find that a predictive-coding model with fewer parameters can match or beat LoF on both the horizon task choices and dream content predictions. If that's the case, it suggests LoF's complexity might be unnecessary – a simpler explanation covers it.

22.4.9 Hall-of-fame and transparency

All accepted red-team reports, along with their code and our reproduced runs, will be archived with DOIs for anyone to examine. The project's online companion site will maintain a “Red-Team Dashboard” showing:

- Open challenges: which challenges are currently live, including any pre-funded bounty amounts and deadlines (if we occasionally sponsor specific questions).
- Resolved attacks: a list of past attempts, noting whether they were wins or losses for LoF, and what fixes or updates were made as a result.
- Scorecard: a summary of how many attempts have been made at each target (e.g. how many tried to break invariance, how many tried rival models, etc.) and how the LoF evidence has shifted over time due to these challenges.

22.4.10 One-page red-team registration (paste and fill)

To lower the barrier, we provide a simple form for registering a challenge. For example:

- Team and contact: _____
- Target tier: T0 / T1 / T2 / T3 / T4 / T5
- Dataset(s) / sites: _____
- Frozen artifacts: container digest _____; commit ____; *schema v*; seeds file hash

- Hypothesis and falsifier: _____
- Primary model / rival model: _____
- Manipulation and utility checks (if applicable): VAS ≥ 10 ; $|\Delta\text{appeal}| \leq 5$. How verified? _____
- Privacy stance: DP applied? yes / no (if no, rationale) _____
- OOS metric and split plan: _____
- Timeline and embargo needs: _____
- Ethics statement (no harm, no re-ID, no contact): _____

22.4.11 Where we go next:

We've built the plan, scaled it up, and even stress-tested it with adversarial challenges. Now, before concluding this professional playbook, Section 22.5 offers a final Research note on how to conclusively judge the Law of Fairness in a full-scale trial. In it, we'll see how to statistically confirm when a life's ledger truly balances out – the capstone for declaring the Law tested.

22.5 Research Notes: Equivalence Testing for L(T)

Goal: Provide a rigorous, audit-ready method to test LoF's neutral-closure prediction at end-of-life using equivalence (instead of difference) tests on the life ledger. In practice, we estimate the ledger with our measured proxy $\bar{L}(t) = \int_0^t HCl(\tau) d\tau$, since the true ledger $L(T) = \int_0^T F(t) dt$ (cumulative felt experience) can't be observed directly. All confirmatory tests in this section are only considered when all channels are open and logged (i.e., $C = \{\text{analgesia, sleep, contact, translation, transport, cash}\}$ during the observation window)—if suffering is unrelieved or communication is cut off, we don't expect neutrality to hold and we wouldn't test it as a LoF confirmation.

22.5.1 What we test (estimands)

Let T be the time of “terminal closure” for a person (the last moment of credible consciousness, or a defined endpoint such as exit from study). We focus on summary measures of HCl in the final window before T (and compare to an earlier baseline window):

- Windowed mean: $HCl_{-\tau}(T) = (1/\tau) \int_{T-\tau}^T HCl(t) dt$. (The average HCl over the last τ before T , expressed in standardized units z .)
- Windowed slope: $s_{-\tau}(T) = \text{slope of } HCl \text{ vs. time over } (T-\tau, T]$, typically estimated via GLS regression or a state-space derivative (units: z/day). This represents the trend: is HCl rising, falling, or flat in that last interval?
- Windowed variance: $\sigma_{-\tau}^2(T) = \text{Var}[HCl(t)] \text{ for } t \in (T-\tau, T]$. (Variance of HCl in the last window, in z^2 units.)
- The LoF signatures expected at end-of-life (if LoF holds) can be phrased as:
- E1: Neutral mean — the mean HCl in the final window is approximately zero (no strong positive or negative tilt).
- E2: Bounded drift — the absolute slope $|s_{-\tau}(T)|$ is very small; no steep last-minute surges or crashes (no “death rally” or plunge — it levels out).
- E3: Variance compression — the variance in HCl narrows in the final window compared to earlier baseline, indicating convergence of experience (highs and lows even out near the end).
- We will analyze these patterns within-person (each person as their own case) and then aggregate evidence across sites/individuals.

22.5.2 Equivalence margins (default preregistered)

We treat the above signatures as *equivalence* hypotheses – we’re testing if mean ≈ 0 , slope ≈ 0 , and variance is *not greater* than some small amount. Based on our preregistered “neutrality” bounds (and what’s practically meaningful), we set margins as:

- Mean Neutrality margin: $\delta_{\text{mean}} = \pm 0.15 z$. (The final-window mean must be within ± 0.15 standard deviations of zero for us to call it neutral.)
- Bounded drift margin: $\delta_{\text{slope}} = 0.05 z/\text{day}$. (We care about the magnitude of slope, so this is effectively ± 0.05 , but since we test $|s|$, a single 0.05 bound on absolute value is used.)
- Variance compression margin: target drop $\kappa = 0.20$ (20% or more reduction in variance). In other words, we expect σ^2 in the final window to be at most 80% of a comparable baseline period. So the non-inferiority margin is $\sigma^2_{\tau(T)} \leq (1 - \kappa) \sigma^2_{\text{base}}$ (i.e., final variance ≤ 0.80 of baseline variance).
- (*If HCl wasn’t standardized to z, we convert it or adjust margins accordingly. Also, sites can tighten these margins if they have power to detect smaller differences – but they should declare that in advance.*)

22.5.3 Windows, baselines, censoring, channels.

- Windows: Preregister one or several τ values for analysis (common choices: 24 h, 72 h, 7 d before T). Each window will be tested separately, with a correction for multiple comparisons if using multiple τ (e.g. Holm’s method).
- Baseline variance: Define a baseline period of the same length earlier. For example, if $\tau = 24$ h, take the window from $T-3\tau$ to $T-2\tau$ (i.e., 2–3 days before T) as the baseline. Compute σ^2_{base} for that window. Ensure channels are also open during baseline if possible, or note differences.
- Censoring: Exclude data after a person has effectively lost consciousness or when heavy sedation starts. In medical contexts, you might define T as the last responsive time, and not include any data after morphine (or other sedation/analgesia) crosses a preregistered threshold, etc. Preregister clear rules (e.g. “if patient on ventilator or RASS ≤ -4 , stop ledger tracking”). This ensures we’re testing the conscious experience period only.
- Channels rule: Report the fraction of time in each final window that all channels were open. Also, decide a minimum threshold for including a case in confirmatory analysis – e.g. “We include a person’s data only if $>70\%$ of the last τ hours had all

“channels open”. If someone’s final days were spent largely with pain unrelieved or isolated, the analysis is non-confirmatory by design. We preregister a “channels open” threshold (e.g., $\geq 70\%$ of minutes with analgesia, sleep, and human contact available) and treat cases below that threshold as non-confirmatory. Poor channel access never “defeats” LoF; it limits what our instruments can confirm.

22.5.4 Models and tests

For frequentist analysis, we will use Two One-Sided Tests (TOST) for equivalence (and non-inferiority for variance):

- E1 Neutral mean (TOST): For each person i , test $H_0: \text{HCl}_{\{\tau,i\}(T)} \leq -0.15$ or $\text{HCl}_{\{\tau,i\}(T)} \geq +0.15$ vs. $H_A: -0.15 < \text{HCl}_{\{\tau,i\}(T)} < +0.15$. We compute $\text{HCl}_{\{\tau,i\}(T)}$ and its standard error (propagating uncertainty from the HCI measurement model). We then perform a TOST procedure (with a small-sample df adjustment if needed) to see if we can reject the extremes. After doing this for each person or site, we can combine results using a random-effects meta-analytic approach (each person provides an effect; we see if the pooled effect is within ± 0.15 with sufficient confidence). Essentially, we want to say “the mean is equivalently zero” within a tight bound.
- E2 Bounded drift (TOST on slope): Similarly, test per person $H_0: |s_{\{\tau,i\}(T)}| \geq 0.05$ vs. $H_A: |s_{\{\tau,i\}(T)}| < 0.05$. Estimate $s_{\{\tau,i\}(T)}$ via either a GLS regression or directly from the state-space model’s slope estimate, including autocorrelation in the error if needed. Apply TOST with the ± 0.05 bounds. Then aggregate evidence across persons (we expect to confidently exclude slopes as large as 0.05 z/day in either direction).
- E3 Variance compression (non-inferiority test): Test per person $H_0: \sigma^2_{\{\tau,i\}(T)} \geq (1 - \kappa) \sigma^2_{\{\text{base},i\}}$ vs. $H_A: \sigma^2_{\{\tau,i\}(T)} < (1 - \kappa) \sigma^2_{\{\text{base},i\}}$, with $\kappa = 0.20$. We can use a variance ratio test or bootstrap. For example, calculate the one-sided 95% CI for the ratio $\sigma^2_{\{\tau,i\}(T)} / \sigma^2_{\{\text{base},i\}}$ (e.g., via a Satterthwaite approximation or resampling that respects autocorrelation). If that entire CI lies below the 0.80 threshold, then we have evidence of compression for that person. We preregister $\kappa = 0.20$ as the target reduction.
- Additionally, we can use mixed-effects models to pool evidence: e.g., fit a random-intercept model $\text{HCl}_{\{i,t\}} = \alpha + u_i + \varepsilon_{\{i,t\}}$ on the final-window data for E1, then check that α (the pooled mean) is within $(-0.15, +0.15)$ with a 90% CI (equivalence). For E2, include time as a fixed effect and random slope if needed and check that the fixed time effect is within ± 0.05 . For E3, do a likelihood-ratio

test comparing a model that forces variance drop $\geq 20\%$ vs. one that doesn't, etc., acknowledging any over-dispersion or AR(1) structure. These are secondary confirmations; the primary method remains person-level tests aggregated.

For Bayesian analysis, we complement this with a Region of Practical Equivalence (ROPE) approach:

- E1: Define ROPE_mean = $[-0.15, +0.15]$ for HCI_{τ} . Compute the posterior for each person's final-window mean (from a state-space model, for instance). We declare it equivalent if $\geq 95\%$ of the posterior mass lies within $[-0.15, +0.15]$. We can also look at the population-level posterior (random-effects mean) and require that to lie in the ROPE too.
- E2: ROPE_slope = $[-0.05, +0.05]$ (z/day) for the slope. Again, require posterior mass concentration within that range.
- E3: Let $\rho = \sigma^2_{\tau(T)} / \sigma^2_{\text{base}}$ be the variance ratio. ROPE_ρ = $[0, 0.80]$ (we're interested only in the upper bound really, but effectively we want the posterior of ρ to be entirely below 0.80).
- If, for each metric, at least 95% of the posterior probability is inside the ROPE (and especially if the group-level effect is inside), we consider that strong evidence of the LoF-predicted neutrality (success).

22.5.5 Power and sample size (practical guidance)

- Per person: Aim to collect enough HCI data in the final window to reliably estimate these measures. A rule of thumb: at least ~ 60 independent HCI points in each final window (for example, if you have EMA 6 times a day plus continuous passives, 10 days might give ~ 60 points). More is better, but often limited by real-world situations.
- Autocorrelation effect: Account for autocorrelation in HCI when determining effective sample size. If ρ_{AR} is the AR(1) autocorrelation, the effective number of observations $n_{\text{eff}} \approx n_{\text{pts}} \cdot (1 - \rho_{\text{AR}})/(1 + \rho_{\text{AR}})$. For instance, with $\rho_{\text{AR}} = 0.5$, you effectively halve your independent sample.
- Across persons: If you consider each person as a study, how many people do you need? Determine this via preregistered power/simulation that reflects your expected within-person precision, autocorrelation, and missingness; plan conservatively given dropouts and missingness.

- Sequential designs: Because end-of-life data collection can be unpredictable, you might use a sequential analysis plan (alpha-spending). For example, use a Pocock boundary or similar to allow an analysis after each batch of say 20 patients. Make sure any sequential plan and alpha adjustments are *pre-specified* and simulation-tested. The margins and models must be frozen up front, and ideally include a simulation notebook in the prereg (perhaps varying autocorrelation to ensure coverage).

22.5.6 Measurement, missingness, sensitivity

- Measurement invariance first: As always, test the HCl measurement model's invariance across any groups *before* pooling data for these equivalence tests. If you have multiple sites or two different devices in the end-of-life data, ensure at least configural and metric invariance hold across them. If metric invariance fails (i.e. different groups interpret HCl differently), then you cannot pool or compare them – restrict to within-person tests and report that cross-person comparisons are not strictly valid.
- State-space uncertainty: Propagate uncertainty from the HCl estimation into your final tests. If using a state-space model, get the posterior or filtered uncertainty of HCl at each time point and incorporate that into the SE of the mean and slope. This might mean using a Kalman filter's error covariance or a bootstrap of the filtering process. The goal is not to pretend we measured HCl perfectly – we carry forward its error.
- Missingness: There will be missing data (e.g. a patient was unresponsive for a few hours, or a sensor failed). Use multiple imputation for sporadic missing questionnaires, and state-space smoothing or interpolation for continuous streams, always carrying forward the uncertainty into the final analysis. If large chunks are missing, consider that akin to channels closed or data not usable for confirmatory analysis.
- Sensitivity analyses: Preregister a small battery of sensitivity checks. For example, if you set $\tau = 24, 72, 168$ hours, test all three. If you had to choose some state-space smoothing parameter, try a “rougher” vs. “smoother” setting. If you set a rule for channel openness (e.g. $\geq 70\%$), maybe test 60% and 80% as bounds. The idea is to see if your conclusions (equivalence or not) hold under slight variations. If they flip only under very specific assumptions, that's a fragility to note.

22.5.7 Negative controls, exclusions, confounds

- Negative controls: Include at least one measure that should not be affected if LoF is true. For instance, a pure motor reaction time series recorded in parallel with HCI (if someone is doing a simple tapping task) should not show any neutrality or compression pattern – if it did, that would imply a general flattening not specific to hedonic experience. Another example: if a “sham” sensor provides random data, it should not exhibit any significant change.
- Exclusions (preregistered): Determine criteria for excluding cases from confirmatory analysis. E.g., if a person’s channels weren’t sufficiently open (they got heavy sedation >50% of the window, or pain was unmanaged), exclude them as a confirmatory case. Also exclude if <60% of the needed data points are valid (huge gaps) or if a device catastrophically failed. Essentially, if the data quality or conditions invalidate the test, don’t count it in confirmatory stats (but still report it in an appendix).
- Confounds to log: Keep a log of potential confounds for each patient: introduction of new opioids, episodes of delirium, mechanical ventilation status, communication restrictions, etc. Many of these relate to channels being closed indirectly, but logging them allows post hoc analysis or at least transparency. They can be recorded as extra fields in the channel log (like “analgesia = 0 due to opioid tolerance” or similar).

22.5.8 Reporting template

To standardize reporting of these equivalence results, we provide a template. An example using a 24 h window might look like:

- Question: Do windowed HCI means approach neutral, with bounded drift and variance compression, before loss of consciousness? (We report the channel-open rate; if channels were substantially closed, the test is classified as non-confirmatory.)
- Design: Observational end-of-life telemetry; latent HCI via state-space model; analyzing windows of $\tau = 24$ h (and 72 h, 7 d) before last consciousness; confirmatory tests via TOST and Bayesian ROPE; random-effects meta-analysis across sites.
- Equivalence margins: Mean ± 0.15 z; slope ± 0.05 z/day; variance $\geq 20\%$ reduction versus baseline.

Results (24 h window; placeholders to fill):

- E1 (mean neutrality) – TOST: [passed/failed]. Pooled $\alpha \hat{=} []$, 90% CI [,], evaluated against ± 0.15 .
 - E2 ($|slope|$ bounded) – TOST: [passed/failed]. Estimated $|s| = []$, 90% CI [,], evaluated against ± 0.05 z/day.
 - E3 (variance compression) – Non-inferiority: [passed/failed]. Variance ratio = []; one-sided 95% CI upper bound = [], evaluated against < 0.80 .
 - Heterogeneity: $I^2 = []\%$. Channel-open rate in this window: []. Negative control streams: [flat / notes].
 - Sensitivity: [summary of preregistered sensitivity checks and whether conclusions changed].
 - Deviations: [none / list].
- Data and code: Deposited ([DOI]); tier-1 individual data in secure enclave (or equivalent).

This concise format hits all key points: what was asked, how it was tested, what the bounds were, the outcomes, any heterogeneity or caveats, and where data can be found.

22.5.9 Failure patterns (predeclared)

Even with equivalence testing, we need to state what patterns would count as LoF failures in end-of-life context. If any of these occur consistently (with channels open and good measurement), they weigh heavily *against* LoF:

- No neutrality at all (E1 fails): If the final-period mean HCl stays outside the ± 0.15 range for *all* reasonable window lengths τ (e.g. consistently positive or consistently negative end-of-life affect beyond the margin). This would mean a strong terminal bias, not neutrality.
- Sustained drift (E2 fails): If $|s| \geq 0.05$ (in absolute z/day) in the final window, indicating a significant upward or downward trend right up to the end (e.g. a person getting exponentially happier or sadder with no leveling). A consistently non-zero drift toward the end violates the “bounded drift” expectation.
- No compression or variance expansion (E3 fails): If the end-of-life variance does not drop below 0.80 of baseline (i.e. ratio ≥ 0.80) or even increases, contrary to expectation that experiences converge. Essentially, if things get more erratic or stay as erratic as ever at the end, LoF’s closure prediction fails.

- Extreme sensitivity: If our conclusions about neutrality/compression flip only under very particular choices (say only one specific window or one particular imputation method), that fragility would be noted. But the more clear-cut Fail pattern is if the above E1–E3 simply do not occur despite ideal conditions.

Each of these would “count against” LoF – especially if observed in multiple independent datasets – and would prompt either a narrowing of LoF’s scope or abandoning the claim for neutral closure.

22.5.10 One-page checklist

Finally, before declaring an end-of-life test complete, run through a quick checklist:

- HCl measurement-model invariance confirmed (configural and metric across any groups).
- All channel fields were logged; channel-open fraction in final window \geq preregistered threshold.
- Windows τ are pre-specified; baseline period defined; sedation/censoring rules applied as planned.
- Equivalence tests set: E1/E2 via TOST (and/or ROPE); E3 via non-inferiority on variance.
- Negative controls showed no spurious “neutrality” or “compression” (flat where expected). Sensitivity analyses run and noted.
- Mixed-effects or meta-analytic pooling done; I^2 reported for heterogeneity.
- Results, code, and simulation notebooks archived with DOIs for transparency.

22.5.11 Where we go next:

With terminal-neutrality bounds and decision rules defined, we now turn to 22.6, where those specifications are put to work in a turnkey, one-week kit. The next subsection packages prereg templates, EMA prompts, a simple dream-valence codebook with blinding instructions, and analysis scripts so a class can test the LoF sleep signature: pairing daily mood drift with next-morning dream affect, running the preregistered model, and reporting pass/fail against the equivalence thresholds set in 22.5.

22.6 Classroom Dream Counterweights

One core prediction of the Law of Fairness is that after especially bad days, people might unconsciously *compensate* with something good – perhaps in their dreams. In other words, our toughest days could be followed by unusually positive or comforting dreams (and exceptionally great days might prompt unsettling dreams), as a way of the mind balancing the “ledger.” This section’s purpose is to let you test that hypothesis in a simple, safe way. You’ll design a small study where participants track their daily mood and their dream emotions, to see if that inverse relationship – a dream counterweight effect – really appears. It’s an accessible experiment to explore whether our brains naturally seek emotional balance overnight.

22.6.1 Study design overview

The experiment is essentially a two-week mood-and-dream journal project. Each participant (it could just be you and your classmates) will rate their mood at the end of each day and then record the emotional tone of their dream the next morning. By collecting this data daily, you create a paired timeline of day vs. night for each person. We’ll then analyze these timelines for patterns: specifically, do we observe that *highly negative days* tend to be followed by *positively toned dreams*, and perhaps that *very positive days* might be followed by *negative or challenging dreams*? The design is straightforward – no special equipment needed, just consistency and honesty in journaling. We will also plan the analysis in advance (that’s the preregistration part) to keep ourselves honest about what we’re looking for.

22.6.2 Daily mood tracking

Each evening, participants should record a simple measure of their overall mood or stress level for that day. This could be as easy as rating the day on a scale from, say, 1 (terrible) to 5 (excellent), or writing a one-word summary of their feelings. The key is to capture how good or bad the day was in a consistent way. For example, you might design a Google Form or a paper diary with a prompt like: “How was your mood today overall?” and some scales or smiley faces. It’s important that everyone does this *before* going to sleep, while the day’s feelings are fresh. These daily mood entries create the “daytime” data that we’ll later compare with the nightly dream data. Remind participants to be as honest as possible – a day might be rated low because of a fight with a friend or high because of a fun event. All such context can be noted, but at minimum we need a numeric or categorical mood rating per day.

22.6.3 Dream journal and morning survey

Each morning, upon waking, participants will rate or describe the main emotional tone of their dream (if they remember one). This is the dream valence part of the data. Just like the mood rating, it can be a simple scale: for instance, 1 = very negative dream (nightmare or sad dream), 3 = neutral or no significant emotion, 5 = very positive dream (happy or comforting). If someone doesn't recall any dream, they can mark that as "No dream recalled" (and treat it as missing for dream-valence analyses unless a different rule is preregistered). It's important this is done first thing in the morning, before the day's events color their memory. Participants can write a brief note about the dream if they want (especially if context might matter), but the crucial piece is rating how positive or negative the dream felt. Over two weeks, this will give each person about 14 data points of day-dream pairs. It's a small sample, but patterns might emerge when we combine data across the class or look at an individual's extremes.

22.6.4 Preregistration and analysis plan

Before you peek at any results, you'll set an analysis plan – this is the preregistration. Essentially, you decide ahead of time how you will test for the dream counterweight effect. For example, you might preregister that you will calculate the association between daytime mood and that night's dream valence within-person (e.g., per-person correlations and a pooled within-person estimate via within-person centering or a mixed-effects model). The prediction (per LoF) is a negative correlation: the worse the day, the better the dream (so a day rated 2 leads to a dream rated 4, etc.). You could also plan a simpler analysis: categorize days as "tough" vs. "okay" vs. "great" and see if the following dreams tend to be "good" vs. "neutral" vs. "bad" accordingly, looking for the inversion pattern (bad day \Rightarrow good dream, good day \Rightarrow bad dream). Write down your intended method: e.g., "We will use a Spearman rank correlation between daily mood and next-morning dream rating. We predict a negative correlation. We will also compare the average dream rating after the lowest 3 days vs. after the highest 3 days for each participant." By committing to this plan, you avoid the temptation to cherry-pick once you see the data. After the plan is set, you can proceed to analyze the collected journal data accordingly.

22.6.5 Expected results and interpretation

What would support the Law of Fairness? If you find that on nights following really bad days participants consistently report more positive dreams (and possibly that really good days are followed by more negative or restless dreams), that's evidence in favor of the balancing hypothesis. For instance, say one student's data shows: on her three worst

days (rated 1 out of 5), her next-morning dream ratings were all 5 (very positive dreams), whereas on her best days (rating 5), the dreams were low (2 or 3, perhaps unpleasant). That pattern – inversion of valence – aligns with LoF’s idea that the system seeks equilibrium. If multiple participants show similar trends, those anecdotes build a case that something non-random is happening. On the other hand, what would **not** support LoF? If dreams simply mirror the day (bad day → bad dream because you went to bed upset, good day → happy dream) or if there’s no clear pattern at all, then the hypothesis doesn’t hold in this small sample. It could mean either the effect isn’t real or it’s too subtle to detect with such a simple method. Either way, remember that an unexpected result is *still valuable*. If there’s no inversion, that suggests perhaps dreams aren’t doing the heavy lifting of mood-balancing – or maybe it occurs only under certain conditions. You would honestly report whatever you find.

22.6.6 Practical tips and ethics

When running a two-week study like this, there are some practical considerations. First, compliance: participants might forget a journal entry here or there. It’s okay – build in friendly reminders (a nightly alarm or a teacher reminding the class each afternoon). If someone misses a day, they shouldn’t try to guess their mood or dream – just leave it blank and continue; consistency is more important than perfection. Second, privacy: mood journals and dream entries can be personal. Assure everyone that individual data won’t be made public. Perhaps assign each participant a code or let them submit their ratings privately (through an online form) so only the aggregator (maybe you or the teacher) sees raw data. This way, no one feels embarrassed about what they share. Third, emotional safety: tracking mood can make people reflect on bad days, and some dreams might uncover sensitive feelings. Make it clear that anyone can skip an entry or withdraw from the study if they become uncomfortable. Also, if a participant reveals something concerning (like very severe sadness), have a plan – e.g., involve a counselor or encourage them to talk to a trusted adult (this is part of ethical duty, as covered in Section 22.8.4). Overall, emphasize that this project is for learning and personal insight; it’s *not* a test or competition. Everyone should approach it with curiosity and honesty, and no one should feel pressured to have certain kinds of dreams or moods. With these tips in mind, your class can safely and successfully complete the project.

22.6.7 Where we go next:

In the next section, we will introduce a “horizon” experiment that mimics having only a few chances left to do something good. It’s a simple game-like task to see if people’s choices change when they think it’s their *last opportunity*. This will further test the Law of Fairness with an active decision-making experiment.

22.7 A Simple Horizon Task

One of the strongest predictions of the Law of Fairness is that as a person nears the “end of the line” – when they feel they have little time or few chances left – their behavior will shift in a telltale way (seeking balance, making reparative choices). Obviously, we can’t simulate an actual end-of-life in the lab for ethical reasons. But we *can* simulate the feeling of a shrinking horizon on a much smaller scale. This section shows you how to create a horizon task: an experiment where some participants believe they’re on their last opportunity to do something, and others believe there’s more to come. By comparing these groups, we test whether the “last chance” triggers the kind of compensatory or balancing behavior LoF would predict.

22.7.1 Design overview

The core of a horizon task is psychological framing. You’ll set up a scenario where participants make a choice (or a series of choices). The key manipulation is what participants believe about future opportunities. For example, imagine a simple game in which a participant can either take a small reward for themselves (an indulgent choice) or give a small benefit to someone else or to a meaningful cause (a reparative or “high-Φ” choice). We present this choice to two groups: Group A is told they will play this game 20 times (so this is just Round 1 of many), whereas Group B is told they will play only once (so this is their only shot). In reality, we might only let Group A actually play a couple of rounds just to uphold the *illusion* of repetition, but the important part is their mindset: “plenty of chances” vs. “now or never.”

22.7.2 The choice

We need a choice that has roughly equal immediate appeal in two directions: one leaning toward personal gratification and one toward compensatory good. It should be simple and relatable for participants. For instance, “You have 5 minutes of free time: would you rather spend it writing a quick thank-you message to someone who helped you recently, or watching a funny YouTube clip?” Both are moderately attractive options to a teenager. One is prosocial/reparative (expressing gratitude, which in LoF terms could help “balance” by adding positivity to someone else’s life and meaning to yours), the other is an indulgent quick pleasure (a laugh now). We design it such that in a vacuum, people might be split on which they prefer – there’s no obvious “right” answer, making it a fair test of the condition effects.

22.7.3 Long-horizon condition

Participants in the long-horizon group (Group A) are told something like: “*You will get to make choices like this on 10 different days this semester.*” This implies if they don’t

choose the thank-you note today, they might choose it another time – plenty of opportunity for acts of kindness later. It's essentially not the last round. (We don't actually need to run all 10 days; we could debrief after one trial. But to keep it believable, you might actually follow up with them over multiple days, or clarify afterward that the multiple rounds were hypothetical – just be clear in debriefing so they aren't confused or feel tricked.)

22.7.4 Short-horizon condition

Participants in the short-horizon group (Group B) hear: "*You will do this only once – this is the only opportunity you'll have to make this kind of choice.*" Now the participant knows it's effectively their last and only chance to either do that thank-you note or watch a funny video. There is no future session where they can "make up" for not doing the kind act. This framing puts a bit of implicit pressure: *now or never*. They believe if they don't choose the generous option this time, there won't be another chance to balance it out later.

22.7.5 random assignment

To avoid bias, participants shouldn't choose their condition. Randomly assign who gets the "one-shot" instructions vs. the "many-shots" instructions. This could be done by flipping a coin for each person, or saying every other name on the class roster goes to condition B, etc. The important part is that it's not based on any personal characteristic (and certainly not self-selected, because maybe kinder people would choose to be in the thank-you scenario!). Random assignment ensures any difference in choices we observe is due to the manipulation – the perceived horizon – and not due to pre-existing differences between groups.

22.7.6 Blinding and concealment

Ideally, participants in Group A and Group B should not compare notes until after the experiment, or they'll figure out the manipulation. If doing this in a single classroom, you could have two versions of the instruction sheet or webpage and ask students *not* to talk during the task. Alternatively, run the two conditions in different classes or at different times. Blinding here is mainly about preventing participants from knowing the true purpose: we might simply tell them we're studying "decision-making preferences" without mentioning fairness or LoF. Those collecting or analyzing the data don't need to be blinded to condition in this simple design (since condition is an objective scenario difference), but it's still good practice that the person analyzing results is not the same person who directly interacted with participants during the task. In a class experiment, for instance, the teacher could gather the choice data without telling the student

analysts which condition is which until after they've done the analysis. There are many small ways to inject some blinding to keep everyone honest.

22.7.7 Data collection

The outcome measure here is straightforward: did the participant choose Option A (e.g. write the thank-you note) or Option B (watch the YouTube clip)? That's a binary outcome for each person. You'll end up with something like, "In the one-shot group, 12 out of 15 chose the thank-you note. In the multi-shot group, 7 out of 15 did." You might also collect a quick self-report after the choice: for example, ask "How satisfied are you with your choice?" or "How meaningful vs. fun was that choice to you?" on a short scale. LoF might predict that those in the one-shot condition who chose indulgence could feel less satisfied afterward (a bit of anticipated regret), but that's a secondary detail – the primary outcome is the choice itself (kind act vs. self-indulgence).

22.7.8 Expected outcome and interpretation

If LoF's idea of horizon-driven balancing is correct, we expect more reparative choices in the short-horizon condition. In our example numbers above, 12 vs. 7 choosing the thank-you note is in that direction. We'd analyze this with a simple comparison (for instance, a χ^2 or Fisher's exact test would not be statistically significant in this hypothetical small sample, but the trend is visible). With a larger sample the difference might be clearer. What's crucial is *why* this would happen: in the last-chance scenario, people might feel an implicit push to do something meaningful or "right," consistent with balancing out any personal gains with a prosocial act. In the plenty-of-chances scenario, that urgency isn't felt, so more people go for immediate fun or at least we'd see a closer mix of choices.

Now, what would not support LoF? If both groups choose the thank-you note at similar rates, or if – surprisingly – the one-shot group is *less* likely to choose the reparative act, that would contradict the LoF hypothesis for this task. It might mean either the effect isn't real or our task didn't capture it (maybe writing a thank-you note isn't universally seen as "balancing" enough, or maybe the scenario wasn't believable). Remember, a null or opposite result is not a failure; it's data. It could tell us that this particular operationalization of "horizon" didn't matter to people, or that other factors (like a personal tendency toward procrastination or skepticism about the setup) overwhelmed the effect. We'd report it and consider what might explain it.

22.7.9 Variations and extensions

The horizon task idea can be adapted in many ways. Instead of a one-off binary choice, you could design a little game where participants accumulate points and have options each round that either give them points or spend points to do something nice (like donate

to a group pot that “benefits everyone” at the end). Then manipulate whether they think the game has many rounds or is about to end. Another twist is measuring not just choices but emotional state: does someone in a last-round condition report different feelings? LoF might predict a kind of focused or wistful mindset knowing it’s the final opportunity – perhaps more reflection or even relief after choosing a reparative act, as if they settled something internal. Those aspects can be explored with surveys after the task to complement the behavioral outcome.

22.7.10 Ethical boundaries

However you design it, keep ethical boundaries in mind. Don’t deceive participants about real consequences. Telling them “this is your only round” when actually you planned only one round anyway is fine (as long as you debrief later that the study was about decision framing). But don’t, for instance, promise a big reward *just* to see if they’ll donate it and then not give it – that would be unethical. In class settings, it’s best to use hypothetical stakes or very minor real stakes (like a token prize or a small amount of extra credit that everyone gets regardless of choice, revealed afterward so no one feels tricked). The focus is on the decision process, not the reward itself, so there’s no need to introduce anything that would cause real distress or disappointment.

In summary, a simple horizon task is one of the most powerful small experiments you can run for LoF. It operationalizes the abstract idea of “end-of-life fairness” into a here-and-now decision. If your results show the predicted tilt toward reparative choices when the horizon shrinks, you’ve demonstrated a core concept of this book in microcosm. If not, you’ve learned something about conditions where the theory might not apply, or about how to improve the setup. Either way, you’ve advanced the inquiry – and hopefully had fun seeing psychology in action.

22.7.11 Where we go next:

With the horizon experiment in hand, you’ve practiced manipulating scenarios to test LoF. Next, we need to ensure such experiments are done responsibly. In 22.8, we shift focus to the ethical guidelines and practical safeguards for student-led research. This includes getting proper consent, protecting privacy, and using blinding where possible. In short, before you run off to test LoF on your peers, we want to make sure you know how to do it in a way that’s fair and safe for everyone involved.

22.8 Ethics and Blinds for Teens

When you run studies in a university lab, an Institutional Review Board (IRB) or ethics committee has your back (and occasionally breathes down your neck) to ensure participants are treated right. In a high school or community setting, you might not have a formal IRB, but the ethical principles are the same. This section is all about doing the right thing when your participants are your peers, classmates, or other young people. We'll cover two big topics: obtaining proper consent (and assent) and using blinding and privacy measures to prevent bias and protect participants. The good news is that these aren't overly complicated – they just require foresight and honesty.

Every research participant has the right to know what they're signing up for and to make a free choice about it. For adults, a consent form and a conversation usually suffice. For minors (typically under 18), you need parental consent and the minor's own agreement (called *assent*). In practice, this means if you're running a study in your school, you'll prepare a simple form for parents to sign that describes the study in plain language and gives them a chance to say "no, I don't want my child in this." At the same time, you'll explain the study to the students themselves (in terms they can understand) and make it clear that participation is optional. Even if Mom or Dad said "yes," the student can decline or withdraw at any time. Participation must be voluntary, not pressured by teachers, grades, or friends.

Some tips for ethical consent with teens: keep the language very clear about what will happen (e.g. "We will ask your child to fill out a 5-minute survey each day for two weeks" or "Your child will take part in a one-hour group activity involving decision-making games"). State the purpose in simple terms ("to learn how people make choices when they have more or less time, as part of a student science project on decision making"). Highlight any potential risks, even if minimal ("They may find some questions about mood or dreams to be personal or upsetting, but they can skip any question they don't want to answer"). Emphasize confidentiality ("No names will be used; we assign ID numbers, and only the teacher will know the ID list"). And explicitly say that it's okay not to participate or to quit early, with no penalty or hard feelings. In a school environment, also clarify that saying "no" or stopping will not affect grades or standing in class at all. We provide a template in 22.9 that covers these points.

22.8.1 Privacy and data handling

Teens (and everyone) have a right to privacy. If students are journaling about their feelings or dreams, those can be very personal. You must take steps to ensure those entries aren't read by the whole class or used as gossip. Strategies include:

- Anonymization: Have students use a code (not their name) on any diary or survey. Only the facilitator (maybe the teacher or the student leader running the study) has the key that links codes to real names, and that key is kept confidential. When discussing data or sharing results, use codes or aggregate information (e.g. “Overall, 5 out of 20 students reported X”) rather than names or identifiable details.
- Controlled access: If you’re collecting data online (say, via Google Forms), restrict access to the responses. The raw data might only be seen by the project lead and perhaps a supervising adult. Classmates who are helping analyze should see only de-identified data. If you have a situation where students are both researchers *and* participants (common in classroom projects), one approach is to split roles: for one project, half the class serves as participants while the other half collects and analyzes data, then they swap roles for a different project. This isn’t always feasible, but the idea is to avoid someone analyzing their best friend’s intimate survey responses – create one extra degree of separation when possible.
- Sensitive topics: If your study touches on anything sensitive (mood, stress, dreams that might reveal trauma, etc.), handle it delicately. Set ground rules from the start: for example, if doing group discussions or activities, establish confidentiality (“what’s shared here stays here”). If journals are collected, consider allowing an “opt-out” or sealed section – meaning if a student wrote something they prefer only the teacher (or no one) reads, they can put that part in a sealed envelope or mark it to be skipped by peer analysts. Always have a plan for if someone’s data indicates they are in distress – for instance, if a student reports very high depression levels or writes something alarming in a journal, the ethical move is to have a policy (known to participants) that a teacher or counselor might be alerted to offer help. Safety trumps privacy if someone could be at risk, but you should be upfront about that limit to confidentiality (e.g. “If you write something that indicates you might harm yourself or others, we will involve the school counselor so you can get support”).

22.8.2 Blinding to hypothesis

We touched on this in the earlier sections: it’s generally good if participants don’t know exactly what you’re testing until afterward. Why? Because humans, especially peers, might alter their behavior if they suspect what result you’re looking for. They might either *try to help* you by doing what they think you expect, or some might playfully *mess with the data*. Neither is good science. With teen participants who might also be your friends, the risk of bias is real – they might want to help you succeed or they might joke around instead

of taking it seriously. So, how to blind? You don't have to lie; you can simply withhold specific details. For instance, in the horizon task, you wouldn't announce "We think if you know it's the last round you'll choose the nice option!" – you'd just present the scenario as a decision-making study without explaining the hypothesis. If someone asks, "What are you trying to find out?" you can give a general answer like "We're studying how different situations might affect choices. I can explain more in detail after we're done." That's truthful enough, and most people will accept it. After the data is collected, you should debrief the participants – tell them what you were looking at and why, and answer their questions. This is important in an educational setting so they learn from the process too. Debriefing also fulfills an ethical obligation: if any deception or withholding of info was used (even mild, like not revealing how many rounds there would be), you must clarify it afterward, explain why it was necessary for the design, and reassure them that it wasn't meant to trick or harm anyone.

Another form of blinding in these studies is blinding the researchers (you and your team) to condition or to participant identities during analysis. For example, when scoring the dream journals, remove names and maybe randomize the order of entries so you don't know which participant or which day a given entry is from while you're rating them. That way you can't (even unconsciously) score entries from "stressful days" more positively, for example. You'd only reunite the pieces (which entry belonged to which condition or person) after all scoring is done. This kind of single-blind or even double-blind setup is standard in professional research and absolutely can be done in a classroom with a bit of organization.

22.8.3 Guarding against coercion

One ethical issue unique to a classroom experiment is the potential for coercion. If a teacher is involved, students might feel they *have* to participate because an authority figure asked, or they fear it could affect their grade. If a motivated student is recruiting peers, those peers might feel pressured socially ("I need to support my friend's project"). It's critical to address this openly. Make it explicit (both in the consent form and verbally) that *not participating will have zero negative consequences*. If the teacher is the one asking for volunteers, it can help to have someone else (another staff member or a student leader not in a grading role) collect the consent forms and surveys, so the teacher doesn't directly know who consented or not (at least until after grades are in). If you're a student researcher, emphasize to your friends: "This is totally up to you – seriously, no worries if you opt out." And mean it: no nagging or guilt-tripping anyone. We want enthusiastic participants or no participants; grudging or fearful participation only muddies the data and violates the spirit of ethical research.

22.8.4 Crisis protocol

This was mentioned in the Part X introduction's non-negotiables: have a plan for if a participant shows distress. In a teen context, that often means if someone starts feeling upset by reflecting on their life or experiences (maybe a survey question unexpectedly triggers a tough emotion, or a discussion about fairness touches a nerve), you stop the activity for that person. Perhaps the student can take a break or be excused without any fuss. If something acute happens – e.g. a participant has a panic attack or reveals something very concerning – you should have a pre-identified adult (school counselor, teacher) to refer to immediately. It's also wise to include on your study materials a note like: "*If you feel upset by any questions, you can stop answering at any time. You can also talk to [Counselor's Name] at the school counseling office if you want to.*" For longer-term or observational studies (like the dream journal over two weeks), periodically check in with participants: "Is everything okay? Remember you can withdraw if you want." It's always better to lose some data than to push someone into discomfort. Ethical research puts the person first, data second.

22.8.5 Special caution for sensitive topics

While we've already advised that students shouldn't be doing actual end-of-life studies or other high-risk projects, it's worth emphasizing: some topics are just not suitable for an untrained researcher in a school project. You might be tempted to do a survey on experiences with loss, severe trauma, or mental health struggles, thinking it relates to LoF balancing somehow. This is generally a no-go for a student project. Those topics can be deeply upsetting and require careful, professional handling. If a student project tries to probe something like that, it should be under direct supervision of a qualified psychologist and with rigorous ethics oversight – which is usually beyond the scope of a class project. Our strong recommendation: steer toward low-risk populations and measures. Dreams, everyday choices, simple mood ratings – these are generally low risk and appropriate for beginners. Leave the high-stakes topics (like serious mental health assessments or anything involving vulnerable individuals in a high-stakes situation) to the professionals until you have the training and approvals to approach them safely.

22.8.6 Debriefing and transparency

After the study, debrief your participants. This means explain what you were testing and what you expect or found. In a class setting, this can be a short discussion or a handout. For example: "*We were interested in whether knowing it's your last chance would change your choice. We had two groups... etc. Our hypothesis was X. We haven't analyzed the data yet, but we will share the results with you once we do. Thank you for helping in this*

experiment!" Debriefing serves two purposes: it's ethical (people deserve to know the true purpose once any deception or withholding is over) and it's educational (participants can learn about the science and see the outcome). If any deception was used (even mild, like misrepresenting how many rounds there would be in the game), you **must** clarify it and explain why it was necessary for the design, and reassure them that it wasn't done to *trick* them for no reason – it was to ensure the results would be unbiased. Being transparent at the end helps maintain trust and shows respect for your participants.

In summary, working with teens and peers requires a bit of extra care, but it's absolutely doable. Get consent from those in charge (parents/guardians), make sure participants themselves are on board without pressure, keep data private, blind what you can, and always have an ethical escape hatch (for them and for you). If you follow these guidelines, your small study will not only yield cleaner data, but everyone involved can feel good about the experience. Science should never come at the cost of someone's dignity or well-being – and if you internalize that now, you're well on your way to being a responsible scientist.

22.8.7 Where we go next:

With ethical foundations laid out, you're prepared to conduct a study the right way. In the next section (22.9), we'll give you practical tools to put these principles into action. You'll find ready-to-use consent form templates and checklists to ensure you've covered things like privacy, data ownership, and opt-out procedures. This final piece will equip you with the paperwork and protocols you need so that when you launch your own LoF experiment, you'll have confidence that it's not just scientifically sound, but ethically sound as well.

22.9 Research Notes: Consent Templates

Obtaining informed consent can sound daunting if you've never written a consent form before. In reality, it's about communicating clearly. In this research note, we provide a template you can adapt for a small-scale LoF study in a school or community setting. This is not legal advice, but rather a starting point for a simple, ethical consent document. Customize the bracketed parts for your project.

22.9.1 Sample consent form template

Consent Form for Participation in Student Research Project

Project Title: [Give your project a name, e.g. "Mood and Dream Journal Project"]

Researchers: [Your name and any team members]

Students at [School Name].

Supervised by [Teacher or Advisor Name].

Purpose of the Study: We are conducting a project as part of our [science/psychology] class to learn about how daily experiences might affect mood and dreams. We hope to understand if patterns predicted by the "Law of Fairness" appear in everyday life.

What You Will Do: If you agree to participate, you will be asked to [describe procedures]. (Examples: "complete a short online survey each evening about your day and each morning about your dreams, for two weeks" or "take part in a one-time 15-minute activity in class where you make some choices in a game, and answer a few questions about your feelings.") All activities will occur during [when/where, e.g. "your free period" or "science class" or "at home via an online form"].

Time Required: The study will take approximately [X minutes per session, over Y days]. (Be upfront: e.g. "5 minutes per day for 14 days (about 70 minutes total)" or "30 minutes on one day.")

Voluntary Participation: Taking part in this study is completely voluntary. You may skip any question or quit the study at any time for any reason. If you choose not to participate or to stop early, there will be no penalty or negative consequences. It will not affect your grades or standing at school in any way.

Possible Risks or Discomfort: We expect minimal risk. You will be asked about ordinary feelings and experiences. However, some questions (for example, about your mood) might be personal or make you reflect on unpleasant feelings. If you feel uncomfortable, you may skip those questions or stop participating. In the unlikely event that participating makes you upset, you can notify [teacher/supervisor name] or talk to the school

counselor [counselor name, contact info]. We have a protocol to ensure anyone distressed gets support.

Benefits: There is no direct benefit to you, except the educational experience of seeing how research is done. The results may help us understand patterns in mood and dreams, and you will be contributing to a class science project. We will share a summary of the overall findings with participants, if interested.

Confidentiality: We will not share your individual responses with anyone outside the project. No names will be used in any analysis or reports. We will assign an ID code to your data. The list linking names to ID codes will be kept by [the teacher/advisor] in a secure manner and will not be revealed to the student researchers until after data collection (if at all). When we discuss results, it will be in aggregate (for example, “X% of students experienced positive dreams after a stressful day”) and no one will be identifiable. If we quote any written response (like a dream description), we will remove any names or specific details.

Data Storage: All data (survey answers, etc.) will be collected via [explain, e.g. “Google Forms accessible only to the researchers” or “paper forms that will be kept in a locked drawer”]. The data will be stored until the project is graded/complete, and then it will be [destroyed or kept anonymized, depending on what you plan]. If we decide to use the data for a science fair or report, it will remain anonymized.

Contacts: If you have questions about the study, you can contact [Your Name] at [email] or [Teacher’s Name] at [email/phone]. If you have concerns about your rights as a participant, you can contact [School administration or an overseeing body, if any, or simply state “the school principal”].

Consent:

I have read the information above (or had it explained to me). I understand that my/my child’s participation is voluntary and that I/we can withdraw at any time. By signing below, I consent to participate (or to allow my child to participate) in this study.

Participant’s:

Name: _____ Signature: _____ Date: _____

If participant is under 18:

Parent/Guardian’s Name: _____ Signature: _____ Date: _____

Assent (for minor participants):

I have had this study explained to me and I agree to take part. I know I can stop at any time. Minor's Signature (if able to provide assent): _____ Date: _____

22.9.2 Notes on the template.

A few notes on the above template: It's written to address both the parent and the student. The parent legally consents for minors, but we also include an assent line for the student to sign, affirming they are on board. In practice, for an older teen (16–17), you can often use one combined form like this. For younger children, assent might be a separate simpler form.

We explicitly mention the Law of Fairness in a brief, accessible way (“patterns predicted by the Law of Fairness”) – this is optional, but being honest about the general aim without jargon is good.

We included who is doing the research (students) and that it's supervised. This builds trust that an adult knows about the project.

Confidentiality is emphasized, because that's usually a big concern. We assure that no names will be used and give a bit of how data will be handled. Notice we also mention if we plan to share results beyond class (science fair, etc.), still anonymized. Always be truthful about where data might go. If it's just a class project and will be discarded, say that. If it might become part of a competition or paper, say that too.

The consent form doubles as a record that the participant (and parent) agreed. Both should sign. If you're doing something online, sometimes a checkbox or electronic signature is used instead – but many schools still prefer paper for parental consent.

We provide contacts – usually an adult or teacher is good to list, since parents might be more comfortable reaching out to them than to a student.

Using the template: You would replace the placeholders with specifics of your study. If your study is the horizon task in class, for instance, the “What You Will Do” would describe the game or choices and the time (e.g. one class period). The risks in that case are almost none, but you still say “no known risks beyond maybe feeling a bit of pressure to choose in the game,” etc. Always mention voluntary nature and no impact on grades. If video or audio will be recorded (probably not in our context, but just in case), you must mention that and how those recordings will be used or destroyed.

Assent for younger kids: If any readers plan to test LoF with younger children (e.g. middle schoolers), you'd prepare a much simpler one-page assent form in kid-friendly language

(and still get parental consent separately). For example: “We are doing a project about feelings. If you want to be in it, we will ask you some questions about your day and your dreams. It will take a few minutes each day. You don’t have to be in the project if you don’t want to. You can stop at any time. It won’t affect your grades. Your answers will be private (we won’t show other kids). Please ask any questions you have. If you want to be in the project, sign your name below.” That could be a child-friendly version accompanying the formal parent consent.

In all cases, keep copies of whatever consent/assent you get. This is your evidence that you followed ethical procedures. Even if you don’t need to show it to anyone (since it’s a class project), it’s good practice.

Finally, remember that a consent form is only as good as the honesty behind it. Don’t downplay risks or exaggerate benefits. If something changes in the study (you decide to add a survey question halfway through), technically you should update the participants/parents and let them re-consent if needed. In a small study, a quick announcement like “We added one question about sleep quality, let us know if anyone has an issue with that” could suffice. Ethics is an ongoing commitment, not a one-time paperwork hoop.

22.9.3 Where we go next:

With the Professional Playbook now complete, we hold every instrument needed to test the Law of Fairness on solid scientific ground — preregistration, invariance checks, open data, replication, and red-team critique. The next step is to assemble all these components into a single, cohesive case. In Part XI — *The Case in One Place*, we leave the level of tools and turn to the argument itself. Here, we integrate the theoretical, empirical, and ethical strands built across the first ten parts to see whether the Law truly holds when every piece is weighed together. We will reconstruct the entire logic chain from raw experience to ledger neutrality, expose every assumption to its hardest objections, and test the Law against full-life data. In short, we move from building the methods to making the case — the decisive synthesis that determines whether fairness is not just a guiding idea but a measurable law of nature.

Part XI — The Case in One Place

For ten parts we've been building the Law of Fairness piece by piece. Now comes the moment to put the entire case together and see if it stands. Part XI assembles everything we've learned into a single argument and defends the Law's guarantee (within its stated bounds). By "guarantee," we don't mean a vague tendency or moral wish—we mean a law-like regularity. In plain terms: for any conscious life that runs its full course, deviations toward misery or toward joy are ultimately counterbalanced, so that the total accumulated experience is exactly neutral at life's close (with any tolerance bands referring only to measurement noise, not to the law itself). This isn't a cosmic reward or punishment, and it isn't something that only happens "on average." It's a strict constraint on what any one stream of experience can do in our world.

To understand what "balance" means here, recall how we measure felt experience throughout this book. We use a composite momentary index (HCl) that merges signals from self-reports, behavior, physiology, and brain activity into one scale of "how it feels." Integrating this index over time gives us an observed ledger of experience. Formally:

$$\hat{L}(t) = \int_0^t HCl(\tau) d\tau,$$

the running total of one's Hedonic Composite Index up to time t . In principle, there's a true (unobserved) ledger $L(T) = \int_0^T F(t) dt$ accumulating the person's actual felt experience $F(t)$ over their entire life $[0, T]$. The Law of Fairness claims that by the terminal closure of a conscious life (what we call the "death of mind"), this ledger is exactly neutral ($L(T)=0$)—neither in surplus of happiness nor in deficit of suffering. We set concrete margins for "neutrality" early on, and every test of the Law uses those preregistered thresholds:

- Neutrality bounds: (i) end-of-life mean within $\pm 0.15 z$ of a matched-baseline neutral point; (ii) terminal slope within $\pm 0.05 z/day$; (iii) terminal variance ratio ≤ 0.80 (see 22.5).
- Compression slope: In the final stretch of life, the trajectory approaches neutrality at a rate within $\pm 0.05 z$ per day. In other words, as the end nears, the average drift flattens toward a zero slope.
- Variance compression: In that same end-of-life window, the variance of experience drops to 80% (or less) of the variance in a comparable earlier baseline window.

When we say fairness is "guaranteed," we mean that—if we measure carefully and respect the limits of our instruments—any lawful trajectory of conscious experience will

satisfy exact neutrality at life's end, with the stated bounds serving only as operational measurement gates for noisy data. All cross-group comparisons in our claims are gated by strict measurement invariance checks (we only compare different people or cultures if we've shown our HCl measure works equivalently for them; otherwise we stick to within-person analyses). And whenever we evaluate models or alternative explanations, we insist on honest predictive performance checks (WAIC, LOO, or log-loss scoring) to avoid cherry-picking comforting fits. In short, every claim about the Law's "balance at closure" comes with guardrails.

How the earlier Parts set the stage for this Part:

Measurement (Parts I–III): We established why no single metric is enough and how a composite index (HCl) can reliably quantify feelings. We learned to fuse surveys with biometrics and brain signals, and to integrate HCl into a ledger $\hat{L}(t)$ with clear uncertainty bounds. These foundations showed that feelings *can* be measured well enough to evaluate a law.

Identity and edge cases (Part V): We defined the unit of accounting not as an immortal "soul" or static self, but as a stream of conscious access. This lets the Law of Fairness apply even when identity is blurred (split-brain patients, dissociative identity, anesthesia gaps) while keeping our claims testable and human-centered.

Evidence for balancing (Parts VI–VII): We identified empirical pillars for the Law. Dreams turned out to be structured "offline" balance adjustments linked to deficits from the day before. End-of-life trajectories showed a telltale compression toward neutrality as time ran out. Long-term studies of life "ledgers" revealed shock-and-recovery patterns consistent with an underlying balance principle. These converging clues are exactly what we'd expect if a true fairness constraint is operating.

Rival explanations (Part VII): We fairly examined leading alternative accounts—predictive processing, hedonic set-points, reinforcement learning/homeostasis models, and hybrids. These rivals can explain stability and homeostatic adjustments in experience, but by themselves none guarantees a neutral lifetime ledger within the strict margins we've set. Where those theories scored points (explaining certain patterns better than LoF), we absorbed their insights as possible mechanisms of the Law rather than as refutations of it.

Ethics and human dignity (Part IX): We made it explicit that no scientific law excuses compassion or action. Even under a true fairness guarantee, "relief is a systems variable"—meaning comfort, dignity, and aid for present suffering always take precedence over

data collection or theoretical predictions. A descriptive law about balance never licenses inaction or indifference in the face of pain.

What this Part will do for you:

- The full case, in one place: Here we consolidate the entire argument for the Law of Fairness—from how we measure feelings to how life might self-correct pain with joy—and present it as one cohesive case.
- Objections answered with evidence: In Chapter 23, we organize the ten toughest objections to LoF into clear “Fail patterns” and answer each with hard data, logical gates, and no wishful thinking. This Part is where we stress-test the Law from every angle.
- Signal vs. script clarity: We separate signal (empirical patterns that must hold if the Law is true) from script (comforting stories or moral narratives we might tell ourselves). The focus stays on what we can observe, measure, and potentially falsify.
- Fairness as a constraint, not a hope: By the end of this Part, you’ll see that measurability, adaptation, prediction, simulation, evolution, and ethics all converge on the same structural conclusion. We appear to live in a constrained system where happiness and suffering cannot drift apart indefinitely—and every life’s ledger closes at neutrality (and therefore must fall within the set fairness bounds when measured, given proper measurement and care)—as operationalized in Parts II and IV.
- Final synthesis and endgame map: Chapter 25 integrates the metaphysical, physical/systems, psychological, spiritual/moral, and societal/ethical lenses into one testable picture. It restates the non-teleological guardrails (a constraint, not a script), enumerates decisive Fail patterns and neutrality gates (equivalence bands, horizon effects, variance compression), and sets a concrete agenda for data, applications, and careful communication going forward.

Chapters in this Part:

- **Chapter 23 — The Ten Hardest Objections (and Our Answers)** - This chapter puts the Law on trial. We state each objection at full strength—“You can’t measure feelings,” “Adaptation explains it,” “You’re moralizing physics,” “Identity is fuzzy,” “Dreams are noise,” “The brain is a prediction machine,” “Simulations prove nothing,” “Evolution wouldn’t select this,” “It’s unfalsifiable,” and “It’s dangerous to say suffering balances”—then answer with decision-grade tests,

explicit Fail patterns, and preregistered gates. Where rivals (adaptation, predictive coding, RL/homeostasis) explain parts of the data, we show what they can and cannot guarantee and how head-to-head comparisons would decide it. The close of the chapter (Research Notes) maps readers to the exact kinds of evidence needed next. The tone is a stress-test—no rhetoric, no moving goalposts—focused on what could truly refute or corroborate LoF.

- **Chapter 24 — If Fairness Is Real** - This capstone distills the discovery claim and its stakes, then turns from debate to action. First, it specifies what “discovery” would mean for LoF (neutral-band closure at the death of mind, supported by the book’s signatures and equivalence tests), and what would count against it. Next, it translates the program for non-specialists—concrete, ethically clean ways ordinary people can participate (e.g., careful self-tracking, dream-ledger exercises, and horizon tasks that respect comfort and dignity). Finally, it calls for courage—not belief—to keep testing the constraint in the open, and closes with a brief reflection on living well under guardrails rather than scripts. The goal is a clear handoff: from Part XI’s stress-test to a shared, humane research agenda if the law’s picture continues to hold.
- **Chapter 25 — Final Synthesis** - This closing chapter pulls the whole case into one frame. It surveys the Law through five lenses (ontological/metaphysical; physical/systems; psychological; spiritual/moral parallels; societal/ethical implications), showing what each lens explains, what it cannot, and how the pieces fit the book’s signatures (neutral-band closure at the death of mind, horizon-driven compression, and equivalence tests). It clarifies that LoF is a systems constraint—not cosmic purpose—reiterates decisive falsifiers and preregistered gates, and translates the synthesis into a forward program for research and practice. The tone is integrative and testable—no platitudes—ending with a call to carry the inquiry forward with rigor and care.

Where we go next:

Now we turn to Chapter 23, which puts the Law of Fairness on trial with the ten hardest objections and decision-grade tests. After the action program in Chapter 24, Chapter 25 closes the Part by integrating mechanisms, measurement, ethics, and communication into a single, testable synthesis.

Chapter 23 — The Ten Hardest Objections (and Our Answers)

Even the strongest theory must survive its strongest critiques. This chapter takes on the most forceful skeptical arguments one by one and shows why none of them overturns the claim that fairness in conscious experience is guaranteed. By “guaranteed,” we mean that for any unified conscious stream that runs to its end, the integrated hedonic ledger of that life will be exactly neutral at terminal closure (with the stated margins referring only to operational measurement gates, not to the law itself). In other words, no life that can be fully observed will end in a gross imbalance of happiness over suffering or vice versa. Mechanisms differ—adaptation in psychology, opponent processes in neuroscience, prediction-error dynamics, care networks among people, even dreams as offline “balance passes”—but the constraint is the same. In our world, an open-ended, ever-worsening imbalance of felt experience just isn’t an admissible trajectory.

We proceed in eleven focused sections (ten objections plus 23.11 Research Notes), each tackling a distinct objection and its corresponding falsifier. Section 23.1 answers the claim “*You can’t measure feelings*” by showing how composite indices (like HCl) do capture subjective experience and how we carry their uncertainty into the ledger $\hat{L}(t)$. Section 23.2 grants that *adaptation* happens (people return toward a baseline after ups and downs) yet demonstrates why mere stability is weaker than true balance; it outlines how we test for over-corrections (compensatory overshoots) rather than just a drift back to baseline. Section 23.3 draws a bright line between describing a law and moralizing about it. It tackles the worry “*You’re just moralizing physics*”, making clear that the Law of Fairness is presented as a physical-style constraint on experience, not a story of anyone “getting what they deserve.” Section 23.4 handles the “*identity is fuzzy*” objection by refining our unit of analysis: it shows that fairness applies to streams of experience, not to some immutable “self,” which addresses cases of split identities or altered states without breaking the testable framework. Section 23.5 considers the skepticism that “*Dreams are noise*” and counters that dreams follow lawful patterns; they serve as structured, experience-balancing mechanisms rather than random nonsense.

Section 23.6 takes on the idea that “*the brain is a prediction machine, so maybe that’s all this is.*” It situates predictive processing as a plausible implementation of the Law’s effects (the brain may reduce surprise or error in ways that incidentally enforce balance), but shows predictive models alone don’t eliminate the need for an overarching constraint like LoF. Section 23.7 addresses “*Simulations prove nothing*”, clarifying that while simulated worlds or agent-based models can illustrate fairness, they’re no substitute for real-world, out-of-sample data. In short, we treat simulations as a way to explore plausibility, not as definitive proof. Section 23.8 considers an evolutionary objection:

“Evolution wouldn’t select for a fairness law.” It explains how evolution, without intending anything, can still yield fairness as a byproduct of viability (organisms that overly favor either extreme happiness or extreme pain would be less fit in the long run), and conversely how a fairness dynamic could enhance resilience. Section 23.9 confronts *“It’s unfalsifiable”* by plainly stating what evidence would break the Law. We specify the exact data patterns that would count as a failure (for example, a final ledger mean far outside $\pm 0.15 z$), and we show how our end-of-life neutrality and compression criteria make the claim empirically testable. Section 23.10 addresses the ethical worry *“It’s dangerous to say suffering balances”*. We emphasize that the Law never licenses indifference to pain—human duties of care remain. This section reinforces how our ethical guidelines (from Part IX) prevent misuse of the idea. Finally, Section 23.11 compiles research notes so you can trace each claim back to where it was operationalized in earlier chapters (for instance, which chapter contained the dream analysis, or the adaptation study, etc.).

Throughout these defenses, we enforce the same four non-negotiable rules:

- Invariance gate: We make cross-group or cross-cultural comparisons only where we’ve shown measurement invariance (configural \rightarrow metric, and scalar when possible). If that condition fails, we restrict the claim to within-person evidence.
- Compression gate: Any statement that a life’s ledger “closes” neutral is always bound by the specific thresholds we set (mean within $\pm 0.15 z$, end-of-life slope within $\pm 0.05 z/day$, and variance ratio ≤ 0.80 vs. baseline in a preregistered end window).
- Model discipline: When comparing Law-of-Fairness models to rival theories, we announce the out-of-sample scoring metric (WAIC, LOO, or log-loss) up front and keep to it. No changing metrics midstream to favor our hypothesis.
- Ethics line: No measurement or argument ever overrides the principle that immediate relief and human dignity come first. If applying the theory would conflict with compassion, compassion wins—every time.

It’s also important to note what this chapter *doesn’t* do. We do not smuggle in any cosmic purpose or moral “desert.” We never claim that the universe *wants* fairness, nor that people *deserve* whatever outcomes they get. We claim only that experience appears constrained in a particular way: across any lawful stream of consciousness, the ledger cannot end in extreme imbalance (given the measurements and gates we’ve set up). If you like a physics analogy, think of it as a kind of conservation law in the domain of

feeling: various processes can convert, buffer, or shift emotional energy around, but by the end of the sequence, the net balance obeys a fixed rule.

We invite you to read these sections with a strict, critical eye. Each objection comes with a clear Fail pattern and an observable falsifier that would invalidate the Law if discovered. If any one of those falsifiers is met in future data (under proper invariance and preregistered conditions), then the Law of Fairness, as stated, is false. As of now, though, the assembled evidence—measurement reliability, dream-driven balancing, opponent-process and prediction-driven rebounds, longitudinal recovery patterns, and end-of-life compression—supports the stronger claim. We therefore defend the Law of Fairness as a true guarantee, not just a tentative tendency.

What you'll get from this Chapter:

- Objections confronted head-on: From “You can’t measure feelings” to “It’s unfalsifiable,” each major critique is presented in its sharpest form and answered with our best evidence and rigorous tests.
- Clarity against rival theories: You’ll see why related ideas (adaptation, predictive coding, reinforcement learning/homeostasis) can all be true and yet still fail to guarantee a neutral lifetime ledger. We outline what distinct evidence would decide between LoF and those rivals.
- Ethical guardrails affirmed: We reiterate why the Law of Fairness never justifies neglecting suffering. The chapter highlights how the ethical principles from Part IX constrain any practical use of the idea and keep our language responsible.
- A real stress-test, not apologetics: This is a genuine attempt to find cracks in the theory. If any objection ultimately holds, it will limit or refute LoF—no moving goalposts. In this chapter we demonstrate that we’re willing to let the Law stand or fall on empirical grounds alone.

Subsections in this Chapter:

- **23.1 “You Can’t Measure Feelings”** - Addresses the claim that subjective experience cannot be quantified. This subsection explains how modern science can measure feelings — from careful self-report scales to neural “signatures” of pain and emotion — and thus how we have the tools to test the Law of Fairness empirically. It lays the foundation by showing that the supposed immeasurability of feelings is no barrier to investigating life’s balance.
- **23.2 “Adaptation Explains It”** - Acknowledges the phenomenon of hedonic adaptation (people returning to baseline after ups or downs) but demonstrates

why simple adaptation isn't sufficient to guarantee the lifetime balance that LoF claims. This subsection outlines how true fairness would entail overshoot and compensation, not just a return to normal. It highlights what patterns — beyond ordinary adaptation — we should see if the Law of Fairness is real.

- **23.3 “You’re Moralizing Physics”** - Draws a clear line between scientific description and moral prescription. This subsection tackles the worry that invoking a “Law of Fairness” smuggles in a moral narrative (as if people get what they deserve). It clarifies that LoF is proposed as a law-like constraint in nature, not a moral judgment, emphasizing that describing a balance in experience is different from claiming some cosmic justice.
- **23.4 “Identity Is Fuzzy”** - Responds to concerns about personal identity and the unit of analysis. It argues that the Law of Fairness applies to a continuous stream of conscious experience, not necessarily to a fixed “self.” By focusing on streams (which can bridge across split-brain cases, altered states, or identity changes), this subsection shows that even if identity is complex or fragmented, each stream still must obey (and can be tested for) the fairness constraint.
- **23.5 “Dreams Are Noise”** - Challenges the notion that dreams and other altered states are just random noise, irrelevant to any balancing process. This subsection counters that dreams exhibit lawful, structured patterns that often compensate for waking experiences. It presents dreams as part of the mind’s balancing toolkit (not mere nonsense), thereby reinforcing that observed dream phenomena support the Law of Fairness rather than contradict it.
- **23.6 “The Brain Is a Prediction Machine”** - Examines the argument that predictive coding alone (the brain minimizing surprise or prediction error) accounts for the patterns we see, potentially making LoF unnecessary. This subsection situates predictive processing as one possible mechanism but shows its limits: by itself, a “prediction machine” brain doesn’t guarantee a neutral lifetime ledger. We learn why an additional fairness constraint (LoF) is needed on top of standard adaptation and prediction processes.
- **23.7 “Simulations Prove Nothing”** - Addresses skepticism about using simulations or thought experiments as evidence. This subsection concedes that while simulations and toy models can illustrate how a fairness law *might* operate, they are no substitute for real-world data. It emphasizes that the Law of Fairness must be supported by empirical evidence, not just by plausible scenarios *in silico*, and outlines the role and limits of simulation in our overall argument.
- **23.8 “Evolution Wouldn’t Select This”** - Considers the evolutionary objection — the idea that natural selection would not favor a built-in fairness law. This subsection explains how a lifetime balance could emerge as a byproduct of evolutionary pressures (for example, extreme imbalances might be maladaptive). It makes the case that evolution doesn’t need to “intend” fairness for a fairness-like constraint to arise, and it shows how such a constraint could actually bolster long-term fitness or resilience.

- **23.9 “It’s Unfalsifiable”** - Directly confronts the critique that the Law of Fairness cannot be tested or disproven. This subsection spells out concrete conditions that would falsify the law (such as finding a life that ends with a wildly non-neutral ledger outside our predefined bounds). By detailing these fail conditions and describing our stringent “neutrality at end-of-life” metrics, it demonstrates that LoF is indeed falsifiable and thus a legitimate scientific claim.
- **23.10 “It’s Dangerous to Say Suffering Balances”** - Tackles the ethical concern that promoting a fairness law could encourage complacency about pain (as if suffering will sort itself out). This subsection reaffirms that the Law of Fairness is a descriptive claim, not a prescription to withhold compassion. It underscores our Part IX ethical guardrails: no matter what balancing tendencies might exist, we must always treat suffering as urgent and real. In short, acknowledging a balance constraint never excuses inaction or indifference.
- **23.11 Research Notes: Where to Find the Evidence** - Compiles a guide to the key evidence underlying each claim in this chapter. This final subsection serves as a roadmap back into the rest of the book: for each objection and answer, it points to the earlier chapters or studies where the supporting data and operational definitions were introduced. Readers get a “source index” so they can trace every major concept (from dream analyses to adaptation studies) back to its origin, ensuring transparency and encouraging further exploration of the evidence.

Where we go next:

Having outlined our approach, we begin with the most fundamental question. Section 23.1 takes on the first and most basic objection — whether we can measure feelings well enough to even test the Law of Fairness — setting the stage for all the challenges that follow.

23.1 “You Can’t Measure Feelings”

Objection: “You can’t measure feelings.” Critics argue that subjective experiences like pain or happiness are inherently immeasurable. Since there’s no ruler or scale for feelings, any talk of balancing suffering with joy is seen as unscientific hand-waving. This objection insists that because feelings are private and qualitative, we cannot quantify them rigorously — and therefore cannot verify any law governing them.

23.1.1 Can feelings be measured?

It’s true that we cannot yet plug a device into someone’s head and read out a single number for “units of suffering.” However, this does not mean feelings are beyond measurement. Science has developed multiple ways to gauge subjective experience. The most direct is careful self-report: people can consistently rank or rate their feelings (e.g. on a pain scale from 1 to 10), and these reports show reliable patterns across individuals. For example, researchers have mapped out a “human feeling space” of 100 core feelings using surveys and brain imaging, revealing a structured, quantifiable landscape of emotion. These findings show that feelings are not ineffable mysteries — they fall into clusters and dimensions (such as positive vs. negative, high arousal vs. low arousal) that can be measured and analyzed scientifically.

23.1.2 Neural signatures of emotion

Moreover, advances in neuroscience show that we can sometimes predict aspects of what someone reports feeling by looking at patterns in brain activity and physiology. A striking example is the development of an fMRI-based “pain signature”: in controlled studies, a specific pattern of brain activation was found to predict an individual’s pain intensity with high accuracy. In other words, by scanning brain activity, scientists could estimate a person’s reported pain intensity — almost like reading an internal thermometer for agony. Similarly, other studies have identified neural correlates for various emotions (like fear or joy) and even created maps of where in the body people sense different feelings.

23.1.3 Testing fairness with measured feelings

Crucially, the fact that we *can* measure feelings (albeit indirectly) means we can test the Law of Fairness. We can compare the “areas under the curve” of someone’s negative experiences and positive experiences over time. If the Law of Fairness holds, those must equilibrate by closure. As discussed in earlier chapters, researchers already use techniques like experience sampling (prompting people throughout the day to report mood) and psychometric questionnaires to compile an individual’s emotional trajectory. Patterns emerging from large datasets show that extreme highs and lows tend to be

temporary, converging toward a mean. This is a first hint of balance. More rigorously, with physiological measures, we could imagine an experiment where participants undergo a calibrated painful experience and then a calibrated pleasurable one, and we check if some balance in neural response or subjective rating is achieved.

23.1.4 Everyday quantification in practice

It's also worth noting that medicine and psychology routinely quantify subjective states. Pain, for instance, is often treated as a "vital sign" assessed by patient report; depression severity is quantified with rating scales. These measurements have predictive power (e.g. a higher pain score correlates with certain stress hormone levels or recovery times). So while feelings are private, they are not unreachable. We can triangulate them through behavior, biology, and self-report.

23.1.5 Ethics and the "forbidden experiment"

On the "forbidden experiment." The maze-rat scenario was kept strictly as a thought experiment to force a clean prediction and then showed how to test the same structure without harm, including precommitment cohorts, hard horizons, and adversarial lures, all implemented in nonsentient or humane tasks. Why keep it at all? Because it prevents vagueness. If, under a hardened horizon, only one path preserves the possibility of later compensation, a true law must concentrate admissible trajectories on that path. That is the standard our proxies must meet — and can falsify. The ethics are nonnegotiable; the science remains testable.

23.1.6 Conclusion: measuring feelings is feasible

The objection that "you can't measure feelings" is outdated. Modern science does measure feelings, both directly (through self-assessment and questionnaires) and indirectly (via brain scans, physiological proxies, and behavioral indicators). These methods have already allowed us to detect the balancing patterns predicted by the Law of Fairness. If anything, continued improvements in neural measurement will make it increasingly possible to verify the fairness balance quantitatively. The subjective nature of feelings is just a challenge that we are steadily overcoming. And as we do, the evidence for a fundamental fairness in the sum of experiences grows ever more tangible.

23.1.7 Where we go next:

The next skeptic argument is that any observed balance is nothing special — just the mind's normal adaptation at work. We tackle this claim by examining how adaptation alone differs from the deeper compensatory balance that the Law of Fairness predicts, moving the debate to whether mere habituation can explain life's emotional ledger.

23.2 “Adaptation Explains It”

Objection: “It’s just adaptation.” This objection accepts that people often bounce back from hardship or lose excitement after windfalls, but attributes it all to psychological adaptation. According to this view, humans (and other animals) simply have built-in mechanisms—like the “hedonic treadmill”—that bring us back to a baseline mood after extreme ups or downs. Thus, any apparent balance of suffering and happiness is not due to a cosmic Law of Fairness, but merely the result of evolutionary adaptation and habituation. In short, the critic says: “No mysterious fairness law is needed; the mind just adjusts to maintain equilibrium, for purely biological reasons.”

23.2.1 Adaptation and baseline mood

It’s certainly true that adaptation plays a major role in our emotional life. We acclimate. A person who wins the lottery may feel euphoria initially, but months later, they often revert to ordinary levels of happiness. Conversely, someone who suffers a serious injury or loss usually finds that their intense grief or pain diminishes over time; they learn to cope and can even find new forms of happiness. Classic studies showed that, one year after their life-changing events, lottery winners and paraplegic accident victims reported similar levels of happiness. This is often cited as evidence of the mind’s resilience and the hedonic treadmill (we have an emotional set-point that we return to). Does this well-understood phenomenon account for the Law of Fairness?

23.2.2 Beyond adaptation: overshoot required

Adaptation is part of the story, but it’s not the whole story. The Law of Fairness goes beyond merely returning to baseline; it implies an overshoot and compensation such that intense suffering is repaid with commensurate positive experience (and vice versa). Adaptation alone would suggest we all hover around a baseline with minor deviations. Fairness, on the other hand, predicts a more complete balancing: e.g. a period of deep suffering might be followed not just by a return to “okay-ness,” but by experiences of heightened joy or meaning that counterbalance the pain. And indeed, we observe phenomena that adaptation alone can’t fully explain.

23.2.3 Opponent processes and relief

One such phenomenon is the “opponent process” dynamics of emotions. Neuroscience shows that when we undergo a strong emotion in one direction, the brain often produces a contrasting after-effect. For example, the cessation of pain can trigger a surge of pleasure and relief—beyond just going back to neutral. Experiments demonstrate that stopping a painful stimulus leads to activation of the brain’s reward centers and a brief euphoric feeling. The intensity of relief can be proportional to the pain that preceded it.

This is more than adaptation to pain; it's an equalization response, almost like the brain is paying you back with a dose of pleasure once the pain ends. Psychologist Richard Solomon described this as the opponent-process theory: every process (pleasurable or painful) is followed by a compensatory opponent process (painful or pleasurable) as the body strives for equilibrium. The dynamics tend to return toward level, and the brain's regulatory processes work to restore that level after deviations.

23.2.4 Dreams as compensation

Adaptation alone typically implies dampening of extremes (habituation). But the Law of Fairness implies a more structured response. Consider dreams of paraplegics from Section 23.5: people who cannot walk in waking life often dream of walking freely. That's not just returning to baseline mood; that's the mind generating a specific positive experience (mobility, freedom) to counterbalance a specific deprivation in reality. Evolutionary adaptation can't fully explain why the brain would bother to do that. It looks like an intrinsic drive to provide fairness in experience, even if only in dreams.

23.2.5 Post-traumatic growth

Furthermore, adaptation usually doesn't predict any overshoot above baseline. Yet people often report post-traumatic growth: profound positive changes because of their suffering. They might find greater appreciation for life, increased empathy, or a new sense of purpose after surviving adversity. Surveys find that over half of trauma survivors report some degree of positive personal growth following the event. In many cases, they don't just bounce back—they bounce forward in certain aspects. While some of this may be reframing or a “silver lining” effect, the consistency of such reports suggests that extreme suffering can lead to equally significant positive transformations. Adaptation alone (which would suggest a return to status quo) doesn't account for these plus-side changes.

23.2.6 Biological homeostasis and balance

It's also useful to consider physiology. The body maintains homeostasis in concrete ways: if you go into a very cold environment (extreme negative stimulus), your body not only shivers to adapt but might later produce a feeling of warmth or numbness as compensation. If you consume a lot of a drug that produces pleasure, the brain adapts by reducing natural dopamine (leading to a crash or pain later). These examples show symmetrical responses to stimuli. They support a rule of balance, of which hedonic adaptation is one manifestation. But the Law of Fairness suggests a principle operating at a higher level of aggregation—across one's life or consciousness, not just in immediate stimulus-response. Adaptation explains the mechanism (how balance can be restored),

but it doesn't explain why the balance is so thorough. Why do these opponent processes exist in the first place? One might argue evolution put them there to keep us motivated (if pain never ended in relief, we'd succumb to despair; if pleasure never waned, we'd stop seeking rewards). That may be true, yet it simply pushes the question one level deeper: why is the optimal evolutionary design one that ends up balancing pleasure and pain so neatly? We could be designed to feel mostly pleasure and just a tiny bit of pain as a warning signal, for instance, but instead nature gave us a system where excess pleasure often carries later costs, and intense pain can be followed by strong relief. This can resemble a built-in balancing dynamic shaped by evolution, without implying any intention or moral plan.

23.2.7 Adaptation vs. true fairness

In summary, adaptation is a real and essential concept—our psyches and bodies do adjust to extremes. However, the Law of Fairness encompasses adaptation but also transcends it. Adaptation explains how balance can occur (through habituation, opponent processes, etc.), but it doesn't fully explain why the balance often appears so proportional and fairness-like. The Law of Fairness posits that the lifetime integral of suffering and positive experience must close exactly at zero, regardless of how compensation is distributed across time. Adaptation alone would only guarantee a return to baseline, not an equal and opposite reward. The evidence (from opponent neural processes to post-traumatic growth and dream compensation) points to something deeper: an overarching principle that ensures balance, using adaptation as one of its tools. Thus, while adaptation contributes to the balancing act, it is not an alternate explanation that makes the Law of Fairness unnecessary—it is, rather, one mechanism by which the law is implemented in our biology.

23.2.8 Where we go next:

Having seen that mere adaptation cannot fully account for a lifetime of balanced experiences, we turn next to a related criticism: the worry that invoking a “Law of Fairness” is just smuggling a moral notion into science. In the following section, we address the charge that we are moralizing physics by proposing this law.

23.3 “You’re Moralizing Physics”

Objection: “You’re moralizing physics.” This objection accuses the Law of Fairness of being a kind of wishful magical thinking—essentially smuggling a moral notion of “justice” into the scientific view of the world. Critics might say, “The universe is cold and indifferent. Any pattern of fairness you think you see is just a human projection, like believing in karma or a ‘just world’ because we want to think people get what they deserve. But physics has no fairness principle—things just happen. By claiming a law ensures suffering is balanced by happiness, aren’t you just moralizing the cosmos? It sounds more like religion or the just-world fallacy than science.”

23.3.1 Not a moral law, just balance

It’s crucial to clarify what the Law of Fairness is not. It is not a moral law handed down by a deity, nor a cosmic judge that rewards virtue and punishes vice. It doesn’t say people “deserve” the happiness or pain they get (no victim-blaming, no moral desert). In fact, the Law of Fairness as we formulate it has nothing to do with one’s actions or “goodness” at all—it’s about experiences balancing out regardless of personal merit. This already separates it from the classic just-world hypothesis, which indeed is a well-known cognitive bias where people irrationally assume that good things happen to good people and bad things happen to bad people. The just-world fallacy is a moralization of events that often leads to blaming victims (“they must have done something to deserve it”). Our theory, by contrast, does not imply anyone’s suffering is deserved; it only suggests that if someone does suffer, nature ensures they will experience an equivalent positive offset at some point (and conversely for extreme pleasure). In other words, it’s about balance, not justice in the ethical sense.

23.3.2 Intuition vs. evidence

Admittedly, the language of “fairness” can evoke moral connotations. Perhaps a better term from a scientific perspective would be “balance of experience” or even a boundary-condition neutrality in the sum of conscious valence. But we chose “Law of Fairness” in part because, to any conscious being, a world where suffering is compensated does feel more fair (in the colloquial sense) than one where it isn’t. We must be careful, however, to distinguish emotional intuition from empirical evidence. Are we only seeing fairness because we yearn for it? This is a valid concern. We guard against it by insisting on data and testable predictions. Throughout this work, we have leaned on neuroscientific findings, psychological studies, and logical argument—rather than appealing to metaphysical justice or spiritual karma—to make the case.

23.3.3 Empirical patterns, not wishful thinking

The patterns we've highlighted (like the opponent-process phenomena, dream compensations, hedonic resetting, etc.) are observational and repeatable, not one-off anecdotes. For example, it's an empirical finding that pain offset activates the same reward circuitry that pleasurable stimuli do. It's an empirical finding that people often have dreams that simulate unmet needs or rehearse threats in benign or altered contexts. These are not moral stories; they're facts about how brains and minds operate. We are inferring a boundary condition from convergent empirical patterns, and we specify explicit falsifiers for that boundary condition. The law we propose happens to align with a human sense of fairness, but that alone doesn't disqualify it. After all, humans feel it's "unfair" if energy or matter just disappear, yet conservation laws hold regardless of our feelings. In our case, we might simply be discovering a conservation law of subjective value: the total "hedonic charge" (pleasure minus pain) in a closed system might net to zero, by law, over time.

23.3.4 Descriptive, not prescriptive

In short, we assert that fairness in experience is a descriptive law, not a prescriptive one. It's not saying what ought to happen in a moral sense; it's saying what does happen as a matter of natural course. This is similar to how biologists might note that "in the long run, traits that don't aid survival tend to disappear." That's not moral, it's just a pattern. Likewise, if we say "in the long run, intense pains tend to be countervailed by intense joys," we're making a factual claim open to verification or falsification.

23.3.5 Guarding against bias

To address the skepticism directly: is it possible we're seeing a pattern that isn't really there because of psychological bias? Humans do have a known bias to assume fairness or karma even without evidence. We combat this by actively seeking out counter-cases. Are there individuals who lived and died in unredeemed agony, with no compensatory happiness? Are there people who coasted through life on a high of pleasure with no downturns? If such cases truly exist, they would challenge our law. The historical and personal record, however, tends to reveal that extreme, prolonged one-sided experiences are rare. Life has a way of curving back. Even the mighty and fortunate encounter suffering (if nothing else, aging and loss), and even the downtrodden find moments of joy or meaning. We cite not just anecdotes but statistical tendencies (e.g., regression to the mean in wellbeing, universal patterns in psychological response). So far, these align with the fairness principle.

23.3.6 Physical law analogies

Another point: physical laws sometimes have surprising “moral” echoes without being moral. The second law of thermodynamics, for instance, implies you can’t get something for nothing (no free lunch in terms of energy) — a kind of stern fairness in nature’s accounting. Yet it’s not moralizing; it’s just physics. The Law of Fairness might be a similar kind of principle applied to the domain of conscious experience. It might even have roots in physics if consciousness is somehow enmeshed with physical law (though that ventures into speculation). The key is, we treat it as a neutral law. It doesn’t care about who you are or what you’ve done; it’s simply an equilibrium that is reached.

23.3.7 No call for complacency

Finally, we must recall that acknowledging a law of balanced experiences is not the same as endorsing complacency or fatalism (that will be addressed in Section 23.10). We are not encouraging anyone to endure suffering on the promise of future joy; we’re investigating if that promise is inherently kept by nature. If the evidence showed an unfair universe (where some individuals could suffer unrequitedly or indulge without consequence), we would accept that. But the evidence suggests otherwise.

23.3.8 Conclusion: pattern, not moral judgment

In conclusion, calling the Law of Fairness “moralizing physics” is a misunderstanding. We are not ascribing intention or value judgment to the universe; we are detecting a pattern. We have taken great care to differentiate this from unfounded beliefs like the just-world fallacy, which involve subjective assignments of desert. The Law of Fairness requires no moral scorekeeping—only a balancing of experiential quantities. It remains an empirical hypothesis: elegant, perhaps emotionally satisfying, but ultimately standing or falling on data. And as the data accumulates, the hypothesis that every joy and every sorrow are counter-weighed gains credibility as a law of nature, not a mere wish.

23.3.9 Where we go next:

Having established that the Law of Fairness is a neutral scientific claim (not a moral wish), we face a deeper philosophical question: what exactly does it mean to “balance” experiences for a person, especially if personal identity is not fixed? In the next section, we confront the objection that “identity is fuzzy,” and explore how fairness could operate when the boundaries of the self are not clear-cut.

23.4 “Identity Is Fuzzy”

Objection: “Identity is fuzzy.” This objection questions who, exactly, is supposed to receive the payback for suffering. If the Law of Fairness guarantees that every conscious being’s pain is balanced by pleasure, what happens when personal identity is not clear-cut? People change over time—“you” in ten years might be a very different person. In extreme thought experiments (teleportation, brain splits, uploaded minds), identity can fork or merge. Critics argue that since personal identity is a fluid, sometimes indeterminate concept, the idea of an accounting of suffering vs. happiness for each person becomes meaningless. “There’s no solid self to balance the books for,” they claim, implying that fairness cannot be a strict law if the entity to which experiences accrue isn’t well-defined.

23.4.1 Streams of experience, not souls

This is a profound philosophical challenge, but it’s one that can be met by carefully defining what the Law of Fairness applies to. The law does not necessarily need to be tied to a permanent, unchanging “soul” or personal identity. Instead, it can be formulated in terms of streams of conscious experience. Each moment of experience, with its degree of pleasantness or unpleasantness, is part of a larger tapestry. Identity, in the traditional sense, might just be a narrative or a convenient fiction for tying those experiences together. Modern philosophy and cognitive science indeed suggest that personal identity is not all-or-nothing; Derek Parfit famously argued that identity is a matter of psychological continuity and can be “fuzzy,” but what truly matters are the connections of memory, personality, and consciousness between moments. If we adopt that view, the Law of Fairness need not attach to a strict ego or soul—it attaches to the continuum of experience that loosely gets labeled a person.

23.4.2 The river of consciousness

Imagine that consciousness is like a river: ever-changing water flowing within banks we call “a person.” The shape of the river might alter (changes in personality, beliefs, etc.), it might split into two branches (as in the split-brain cases where one brain produces two semi-independent streams of consciousness), or two tributaries might join (perhaps in a hypothetical future where minds could merge). If fairness is a law, it would apply to the water, not the river’s name. In other words, every unit of “water” (experience) that goes down a painful path must eventually traverse a pleasurable path as well, though it might not be under the same name or even in the same segment of the stream.

23.4.3 Continuity within one life

This may sound abstract, so let's ground it: In ordinary life, even though our identity is somewhat fluid across time, we still treat ourselves as the same person who was sad yesterday and happy today. The fairness law operates at that level—over the span of a life, even though the child you were and the adult you are share few molecules or memories, we consider that one continuum of consciousness and see balance achieved within it. If a person's identity diverges (say, in dissociative identity disorder, where different alters have different experiences), the law would suggest each conscious center still gets its share of balance. If identities merge (imagine a sci-fi scenario of mind fusion), then the combined continuum's experiences would balance out in total.

23.4.4 Death and the scope of balance

What if a person dies after great suffering—when and to whom does the compensatory happiness happen? This is admittedly more speculative. If one does not assume an afterlife (we have not in this scientific context), then perhaps the balancing can occur during life via various avenues (psychological or physical phenomena) and if it seemingly doesn't, it raises a question. Under the stated scope of this book, we do not broaden the unit beyond the unified stream of conscious access: if careful end-of-life measurement still shows a gross, unrectified deficit (under our preregistered gates), that counts as evidence against the Law rather than a reason to shift the scope.

23.4.5 Fairness in the web of consciousness

A more nuanced reconciliation is this: the Law of Fairness is formulated in terms of streams of conscious access, so even when identity boundaries are fuzzy, the accounting remains within that stream. But in edge cases where identity is disrupted, the law doesn't break—because it was never about moral desert of a person, only about matching amounts of suffering and joy in the fabric of experience.

Think of a simpler analogy: conservation of energy in physics doesn't care if you label one part of a system "object A" and another "object B"; if A loses energy, the energy will go to B or C. Similarly, conservation analogies are about labels within a closed system; in this book, the closed system is the unified stream of conscious access we can test, not a cross-person transfer of "payback."

23.4.6 A balance of experiences

To keep it practical: even if identity is fuzzy, experiences are real. Each experience has a value (pain or pleasure magnitude). The Law of Fairness, as stated here, is not a cross-person aggregation claim; it is a closure claim about a unified stream of conscious

access. We already see that within one person's life, the ups and downs often counterweigh each other. That is enough to demonstrate the principle. If identity were truly fragmented, as long as we track the experiential "currency" itself, fairness can still hold.

23.4.7 Splits, merges, and copies

Philosophically, one might ask: if a person splits into two, does each get half the karmic balance? Parfit's view would say personal identity is not what matters; what matters is that the experiences still occur in each branch, and each branch presumably would carry forward its own need for balance. Perhaps each "copy" of the person then has to have its own suffering and joy balanced. If two people merge, perhaps their combined experiences find a new equilibrium together. These are far-out scenarios, but interestingly, none logically refute the possibility of a balancing law—they just complicate to whom we attribute the balance. Our stance is that we attribute it to experience itself, not strictly to a named individual.

23.4.8 Fairness over a changing self

In everyday terms: you are not exactly the same "you" who felt heartbreak at age 16, but you remember it, and perhaps the profound love you felt at 30 helped heal that wound, leaving you with a sense that life gave back happiness after early pain. That's fairness at work, despite the fuzziness of identity over 14 years. The narrative of self provides continuity enough for us to talk about fairness in one life. And if selves can sometimes blur at edges, the fairness principle likely operates at whatever level consciousness maintains continuity.

23.4.9 Conclusion: Experience over identity

In conclusion, the objection that "identity is fuzzy" does not negate the Law of Fairness—it just forces us to refine our thinking. The law does not require a perfectly defined self; it only requires that conscious experiences are part of an interconnected continuum. Fairness can be thought of as a conservation law for the total hedonic content in that continuum. Personal identity is an emergent, sometimes blurry concept, but that doesn't stop the balancing of experiences from occurring. Just as physics can conserve quantities across diffuse fields and entangled particles, the total hedonic ledger is conserved across the relevant stream of conscious access, even if personal identity is fluid. Put simply, even if the "I" who suffers is not rigidly the "I" who rejoices later, what truly matters is that the suffering and rejoicing happen, and they balance each other out in the grand tally of sentient experience.

23.4.10 Where we go next:

If fairness can operate across the fluid boundaries of personal identity, we might ask where evidence for such balancing is most striking. Some of the strongest hints of an intrinsic balancing mechanism come from a surprising realm: our dreams. In the next section, we address the objection that dreams are meaningless noise, and we show how dreams actually provide evidence of the mind working to balance experiences.

23.5 “Dreams Are Noise”

Objection: “Dreams are just noise.” Skeptics here challenge any evidence for the Law of Fairness that comes from dreams or other altered states. They argue that dreams are random byproducts of the sleeping brain—meaningless firings with no adaptive function or patterned content. The classic neuroscientific view (the activation-synthesis hypothesis) holds that dreams result from the cortex trying to make sense of random signals from the brainstem. If dreams are essentially gibberish, then pointing to dreams as a realm where suffering is balanced by satisfying experiences (or vice versa) is unconvincing. “You can’t draw conclusions from dreamland,” they say. “Any apparent fairness in dreams is coincidental or wish fulfillment, not evidence of a law.”

23.5.1 Dreams are not pure chaos

It’s understandable why one might dismiss dreams. For a long time, science did lean toward the idea that dreams are mostly chaotic, a mental static with maybe some leftover daytime thoughts stirred in. However, more recent research has upended the notion that dreams are pure noise. While dreams can certainly be bizarre and don’t follow waking logic, they are far from random in many cases. Emotional concerns, desires, and unresolved daytime experiences often shape dream content in systematic ways. In other words, dreams have structure and purpose. We now have good evidence that dreams serve functions like emotional regulation, memory consolidation, and threat simulation. They frequently reflect a person’s waking life challenges and attempt to process or balance them in some fashion.

23.5.2 Dreams compensate for deficits

One of the most compelling sets of findings in this context involves dreams compensating for deficiencies or traumas in waking life. Consider again the example of paraplegic individuals: Studies have shown that people with paraplegia (even those born without the ability to walk) frequently dream of walking, running, and dancing. In a study of paraplegic dream reports, paraplegic patients experienced walking or other leg movements in roughly 38% of their dreams—about the same frequency as able-bodied people dreaming of such movements. They were rarely disabled within their dreams. This means the dreaming brain was giving them experiences of mobility that they lack when awake. That is an extraordinary form of internal compensation. It’s as if the brain says, “You can’t walk in reality, so you will walk in your dreams.” This is far from random noise; it’s targeted fulfillment of a specific deficit, akin to a fairness-driven response. The subjects who had never walked in life still walked in their dreams, suggesting some innate program or desire was being satisfied during sleep.

23.5.3 Senses restored and traumas resolved

Similarly, people who are deaf from birth often have dreams rich in visual and tactile content, and those who lose their sight later in life continue to see in dreams for years thereafter, maintaining an internal visual world even as their waking world goes dark. Again, the dream is providing something that waking life no longer can. Trauma survivors might experience repetitive nightmares initially (reflecting their trauma), but over time, some dreams begin to integrate or resolve the trauma—either by providing scenarios where the threat is overcome or by placing the person in safe, healing contexts. There is evidence suggesting that dreams help work through negative emotions by linking them with neutral or positive contexts, effectively diminishing the emotional charge of traumatic memories. This is sometimes called a virtual form of exposure therapy that the brain conducts on itself.

23.5.4 The nightly equilibrium workshop

None of this fits the idea of dreams as meaningless noise. Rather, it portrays dreams as a nightly workshop of the psyche, often aiming to restore equilibrium. If you’re extremely socially isolated, you might dream of friends or loved ones; if you’re starving, you’ll dream of food (famously, people in concentration camps during WWII reported dreams of lavish feasts). Freud’s old notion of “wish fulfillment” in dreams was overstated in specifics, but it contained a kernel of truth: dreams do skew toward giving us emotionally important experiences—sometimes wishes, sometimes fears to practice coping with (the “threat simulation” theory). In either case, there is a tilt toward balance: too much fear in life, and your dreams might repeatedly expose you to that fear until you conquer it (thus reducing net fear); too little pleasure or agency in life, and dreams might grant you those very experiences to keep the psyche from withering.

23.5.5 Hints of an underlying fairness

One could argue these balancing dream phenomena are part of adaptation (from Section 23.2) or emotional regulation, which they are. But the salient point is that the brain has a built-in method to deliver experiences when reality won’t. That’s a powerful hint of an underlying fairness principle. The dream world, unconstrained by physical limits, becomes a canvas where the imbalance of waking life can be redressed. It’s telling that these dream compensations often involve core needs: mobility, social belonging, safety, achievement. The things life denies, dreams often provide symbolically or literally. This isn’t guaranteed every single night, of course—many dreams are mundane or fragmentary. But across the tapestry of one’s dream life, patterns emerge. They are not

random TV static; they are more like an ongoing story co-written by our conscious and unconscious mind to make sense of and equalize our experiences.

23.5.6 Dreams are meaningful, not random

To address the objection head on: No, dreams are not mere noise. Even the activation-synthesis proponents now acknowledge that the dreaming brain's "random" activation is modulated by memory networks and emotional salience. The result is that dreams disproportionately feature things that matter to us. Their bizarreness is often just a disguise over deep coherence related to our concerns. Therefore, when we cite dreams as evidence (for instance, the paraplegic dream study or the statistical content analyses of dreams across cultures), we are not grasping at straws. We're using a legitimate data source about the mind's inner workings.

23.5.7 Conclusion: dreams showcase fairness

In sum, the existence of structured dream compensation strengthens the case for the Law of Fairness. It shows the mind actively generates experiences to fill gaps left by reality. If someone were to insist dreams have no meaning whatsoever, they would have to ignore a substantial body of research to the contrary. Dreams might not obey the logic of waking physics, but they do obey a certain psychological logic—one that often points toward restoring equilibrium. So rather than being a challenge to our thesis, the nature of dreams is actually a showcase for it: a clear example where, freed from external constraints, the system creates fairness by night to complement the inequities of the day.

23.5.8 Where we go next:

Having seen how the sleeping brain compensates for waking imbalances, we turn to a different explanation skeptics offer for life's apparent balance. The next objection asks whether the brain's constant prediction and regulation of experience, rather than any fairness law, could account for the patterns we see.

23.6 “The Brain Is a Prediction Machine”

Objection: “The brain is a prediction machine.” This objection suggests that all the phenomena we attribute to the Law of Fairness might instead be explained by the brain’s well-documented habit of predicting and regulating sensory input. The brain constantly generates expectations about the world and adjusts its perceptions and reactions when those expectations are wrong. Critics propose that what we call “balancing suffering with joy” could just be the brain minimizing surprise: after a bad event, the brain predicts (and thus orchestrates) a swing back to good, not because of a cosmic law, but as a way to maintain stability and reduce prediction error. In short, any observed balance might be a byproduct of the brain’s efforts to keep its model of the world consistent, not an external law of fairness.

23.6.1 Predictive processing explained

It’s certainly true that predictive processing is a key function of the brain. Neuroscience models (like those by Karl Friston, Andy Clark, and others) describe the brain as fundamentally engaged in guessing what will happen next and updating its model when guesses don’t match reality. The brain is, in a catchy phrase, a “prediction machine” constantly trying to minimize surprise. How might this relate to our discussion of fairness? One might argue, for example, that if you’re in pain for a long time, your brain might start to expect relief or improvement (since perpetual pain is uncommon in its learned model). That expectation could lead you to notice or even create positive changes (through behavior or endogenous chemicals) that bring relief, thus confirming the prediction. Similarly, after a prolonged pleasure, perhaps the brain predicts a downturn (because it has learned that highs don’t last), leading to self-fulfilling disappointment or a comedown engineered by, say, opponent neural processes. In this view, the brain’s need for predictable equilibrium could mimic a fairness effect.

23.6.2 The brain predicts fairness because it occurs

However, there are several points to consider. First, predictive processing doesn’t work in a vacuum; it’s guided by actual past experience. If life truly had no tendency to balance out extremes, the brain wouldn’t predict a balance—it would predict whatever pattern it had observed. The very fact that the brain often does expect things to even out (e.g. “What goes up must come down” as a learned heuristic) suggests that the environment of experience it has grown up in tends to have that property. In other words, if the brain’s predictions of a rebound often come true, that’s evidence for the underlying fairness dynamic, not an alternative to it. The brain could just as easily develop a pessimistic model (“sometimes things just keep getting worse”) or an eternally optimistic one (“good times will just keep rolling”). But by and large, people intuitively expect balances—

sickness followed by health, stress followed by calm, etc.—because that is what they have repeatedly encountered.

23.6.3 Prediction as fairness mechanism

Second, even if the brain's predictive algorithms contribute to maintaining emotional equilibrium, that could be the mechanism by which fairness is enforced rather than an unrelated coincidence. Perhaps the brain's prediction circuitry is itself tuned to ensure no experience veers too far without correction. It's not mutually exclusive with the Law of Fairness; it could be the implementation of it. For instance, suppose on an unconscious level the brain tracks cumulative error signals related to well-being: too much deviation in the pain or pleasure direction rings alarm bells (surprise), so it initiates counter-regulatory measures (like releasing stress hormones or endorphins, adjusting motivation, etc.) to course-correct. That aligns with fairness—and frames it as an intrinsic principle of brain function.

23.6.4 Prediction vs. feeling content

Another angle: The predictive brain theory mainly addresses perception and action, not the intrinsic generation of feelings. It explains phenomena like why we see optical illusions or why we might not notice something unanticipated. But the Law of Fairness is about the content of experience (pleasant vs. unpleasant) over time. Predictive coding might explain some specific instances, like placebo effects (the expectation of pain relief triggers real relief) or nocebo (expecting a negative outcome produces distress). Indeed, placebo analgesia — you believe you got a painkiller, so your pain diminishes — can be seen as a predictive brain-induced balancing (it expected less pain, so the brain delivered less pain). That's fascinating and indeed supportive of the idea that the brain strives to match pain to what it "should" be. But note, the brain's "should" here was influenced by the belief of relief, an external suggestion. In absence of such suggestion, does the brain spontaneously create relief after pain just to satisfy a prediction?

23.6.5 Learned optimism vs. pessimism

Sometimes yes: if you strongly believe "things will get better soon," that optimism can itself create a self-fulfilling prophecy through improved stress response and resilience. But where did that belief come from? Often from experience—people have learned that most pains do subside with time (because they do, empirically). Our predictive brains incorporate that empirical regularity. If the world were truly cruel and unfair, a rational predictive brain might conclude "if I'm in pain now, it's likely to continue or get worse" and then perhaps do nothing to mitigate it (or even amplify it out of despair). Some depressed individuals indeed fall into such traps: they predict things won't improve, and

this can hamper the brain's balancing efforts, leading to prolonged suffering. That's a case where prediction (albeit possibly distorted) can delay or impede the natural reversion to baseline happiness, resulting in pathology. We intervene therapeutically to break that negative prediction cycle, allowing the person's mood to recover. Again, this interplay suggests the brain's prediction system and the fairness/homeostatic mechanisms are intertwined.

23.6.6 When balance defies prediction

It's also worth pointing out that not all balancing events are easily explained by prediction. For example, consider spontaneous euphoria that sometimes comes after a period of despair — not because anything particular changed externally, but almost as if the brain hit a tipping point and said "enough gloom, time for a swing upward." People with bipolar disorder experience extreme versions of this (depressive phases giving way to manic phases and vice versa) without external triggers. One could describe bipolar as a dysregulated fairness oscillator: it overshoots on both sides. Predictive processing theory doesn't straightforwardly account for why a brain would generate a manic high after a depressed low. In fact, mania often introduces more surprise and prediction error (the world doesn't actually conform to the grandiose expectations). Yet it happens, suggesting an internal push to the opposite pole that's not purely about minimizing surprise. Instead, it looks like an exaggerated enforcement of balance (too much negative, now too much positive). In stable individuals, the same dynamic might be present but damped to a healthy degree — after a rough week you might spontaneously have a day of lightness or creative energy, even if nothing external improved markedly.

23.6.7 Stability Is not enough

In summary, the predictive brain framework is not at odds with the Law of Fairness; if anything, it complements it by describing one way the brain could achieve balance. But it doesn't remove the need for the law, because it doesn't inherently require that pleasure equals pain in sum — it could just aim for stability around a set-point. The Law of Fairness implies something stronger: a compensatory symmetry, not just stability. The predictive model might explain stability (preventing runaway feedback loops of pain or pleasure because that would be surprising against our prior experiences), but it doesn't by itself guarantee symmetry — unless one assumes that symmetry is in the prior experience.

23.6.8 Brain expects fairness because it's real

Ultimately, saying "the brain tries to minimize surprise" is a very general statement, whereas saying "everyone's joys and sorrows balance out" is a very specific one. The latter is not a necessary consequence of the former without empirical support. We have

provided empirical support throughout this book. The brain's predictive nature likely facilitates fairness by pushing expectations toward equilibrium and making it so. Yet, it had to learn that expectation from something real. That something, we argue, is the underlying fairness of experience distribution.

Therefore, invoking the predictive brain does not explain away the Law of Fairness; it likely operationalizes it. We still observe the fairness outcome and must account for why the brain's predictions align with that outcome so consistently. The simplest explanation is that the outcome is a genuine feature of our reality. The brain is a prediction machine, yes, and what it has learned to predict is a world where pain is followed by relief and gain is followed by loss until an equilibrium asserts itself. In other words, the prediction machine analogy doesn't refute the fairness law—it actually hints at it by showing our brains expect fairness and often act to realize those expectations.

23.6.9 Where we go next:

The idea that the brain itself works to enforce balance naturally leads to the question of how we investigate such mechanisms. One way scientists explore the implications of the Law of Fairness is through simulations. In the next section, we address the objection that simulations "prove nothing," clarifying how simulations are used and what they can (and cannot) tell us about real-world fairness.

23.7 “Simulations Prove Nothing”

Objection: “Simulations prove nothing.” Here, skeptics target any computational or theoretical simulations used to illustrate the Law of Fairness. Perhaps earlier in the book, we or others have presented a computer model, thought experiment, or simplified simulation of conscious agents that exhibit balancing behavior. The critic retorts: “So what if you can program a simulation to balance pain and pleasure? That’s just a toy model. Reality is not obliged to follow your simulation. Unless you have direct empirical proof in the real world, simulations are just speculation in digital form.” In short, they caution us not to confuse model results with evidence of actual natural law.

23.7.1 Simulation bias and assumptions

This is a fair reminder—there’s a reason we call them simulations. They are only as good as their assumptions and design. If we made a computer model of mini-conscious entities that invariably compensated each other’s suffering with happiness, we would indeed have built fairness in rather than discovered it. A simulation can definitely be biased by the programmer’s expectations. For instance, if we assume in the code that any pain triggers a healing process that yields pleasure, then of course the simulated data will show balance; but that would be a circular demonstration, not a proof of nature’s law.

23.7.2 Simulations as useful tools

However, simulations can still be very useful in science when used appropriately. They allow us to test ideas in a controlled, repeatable way and to explore the logical consequences of our hypotheses. The key is to not place more weight on them than they deserve. We use simulations to sharpen our intuition, not to serve as final proof.

23.7.3 When simulations align with reality

Consider an analogy: Simulations of planetary orbits (using Newton’s laws) aren’t what proved gravity exists, but they did verify that given Newton’s assumptions, the model output matches observed reality (e.g., the orbits of planets). That concordance between simulation and observation builds confidence in the theory. In our case, if we build a simulation of, say, a society of agents that learn and adapt (with neural-network-like models) and we do not explicitly program a fairness law, yet we find that over time the agents’ experiences still balance out, that’s intriguing. It suggests that fairness might emerge from more fundamental assumptions. If repeated variations of such simulations produce the same result, one might suspect an underlying principle at work, analogous to how many different simulated solar systems all obey conservation of momentum.

23.7.4 Learning from simulation outcomes

On the other hand, if simulations require fine-tuning to get fairness, that teaches us something too: maybe the law isn't as inevitable as we thought under generic conditions, meaning if reality shows fairness, it points to a specific mechanism making it so. Either way, simulation outcomes feed back into hypothesis refinement.

23.7.5 A simple emotional simulation

Let's ground this: Suppose we simulate a very simple model of emotional dynamics. Each agent has a mood that goes up or down based on random life events, but also has internal adaptive dynamics (like a virtual hormone system that tries to keep mood near a baseline). We run this simulation for many agents over long periods. We might observe that indeed each agent's mood graph oscillates around baseline and that extreme deviations self-correct — a form of fairness (no one stays extremely high or low indefinitely). Is this surprising? Probably not, because we explicitly included an adaptive mechanism. It shows the mechanism can enforce balance, but we already assumed it.

23.7.6 Social network simulation

Now suppose we leave out the explicit adaptive mechanism and let agents influence each other in a network (maybe representing social support or contagion of emotions). Will the network as a whole exhibit a balance? Possibly, for instance if one agent's downturn tends to be mitigated by help from others (who themselves feel good for helping, etc.), you might see a community-level regulation. But if we see some agents acting as "sinks" of suffering (absorbing others' pain without relief), that would break fairness locally. By adjusting parameters and rules, we might identify conditions that lead to overall fairness equilibrium (like sufficiently reciprocal support). That's useful: it indicates what real-life conditions might be required for fairness at a social level (e.g., strong social bonds).

23.7.7 Biological simulation

Another scenario: simulation of a physical or biological process. Perhaps a theoretical model of brain homeostasis might simulate neurons and neurotransmitters under stress vs. relief conditions. If the emergent outcome is a balancing of certain signals correlating with pain/pleasure, it strengthens our understanding that the brain could implement fairness spontaneously.

23.7.8 Reality vs. simulation

But the skeptic's core message is correct: alone, a simulation doesn't prove reality does the same. We must always compare the model's output to empirical data. In our work,

we haven't relied on simulations as primary evidence; they have been illustrations. For example, when we mentioned a "thought experiment" or a simplified model demonstrating fairness, it was to help conceptualize how the law might operate, not to claim that because our simulated creatures find balance, humans must too.

23.7.9 Avoiding built-in bias

We acknowledge that if someone programs their biases in, they'll get them out. Therefore, any simulation we use is careful to only incorporate well-established processes (like adaptation, or known physiological feedback loops), and then we check if fairness comes out as a consequence. If it does, that's a consistency check: it shows our hypothesis isn't internally contradictory. If it didn't, that would actually be more concerning (it would mean our understanding of those processes doesn't naturally yield fairness).

23.7.10 Simulations show plausibility

In plain terms: Simulations are supportive tools, not the foundation of our case. They "prove" nothing on their own, but they can demonstrate plausibility. For a contentious idea like the Law of Fairness, showing that it can arise in a model at least dispels the notion that it's impossible or logically incoherent. The heavy lifting still has to be done by real-world evidence—psychological studies, neuroscience, cross-cultural data, longitudinal observations—all of which we have drawn upon in this book.

23.7.11 A balanced perspective

So we agree with the skeptic to an extent: one should not get overly excited because a computer program shows balance in a virtual environment. But neither should one dismiss simulations outright as pointless. They are a means of hypothesis testing under controlled conditions. When used properly, they strengthen the overall argument by demonstrating that our proposed mechanisms (like neural opponent processes, or social support dynamics, etc.) indeed can produce the hypothesized outcome.

23.7.12 Conclusion: simulations as rehearsals

In conclusion, while simulations prove nothing in isolation, ours have served to illustrate and verify that known mechanisms could account for the Law of Fairness. Ultimately, the reason we believe in the law is not because of pixels on a screen or numbers in a log file—it's because those align with patterns we see in real life, as documented in the empirical sources cited throughout. The simulations are the dress rehearsal; reality is the show, and in reality we find that suffering and joy dance in tandem much as our models predicted.

23.7.13 Where we go next:

Even though simulations themselves are not conclusive evidence, they help guide us toward real-world phenomena to investigate. The next objection comes from the realm of evolution: if balancing experiences is real, why would natural selection favor it? In the following section, we tackle the claim that “evolution wouldn’t select this,” examining whether a fairness law is compatible with Darwinian theory.

23.8 “Evolution Wouldn’t Select This”

Objection: “Evolution wouldn’t select this.” This objection comes from a Darwinian perspective: if the Law of Fairness were real, it implies a lot of what happens to an organism (in terms of experiences) doesn’t directly aid its survival or reproduction—since it’s just about balancing a “ledger” of pain and pleasure. Why would evolution favor organisms that internally compensate suffering with happiness? From a strict survival standpoint, pain is a useful signal (to avoid harm) and pleasure is a reward (to reinforce beneficial behaviors). But balancing them for fairness’s sake seems wasteful or even counterproductive. For example, an animal in pain that ‘magically’ gets a compensatory pleasure might lose the drive to escape the cause of pain. Or one that experiences a crash after too much pleasure might forego opportunities. Critics claim that natural selection, being ruthlessly pragmatic, wouldn’t maintain a mechanism whose sole function is to ensure moral fairness. So, they argue, the Law of Fairness is biologically implausible.

23.8.1 Nature’s cruel appearances

This is a thought-provoking critique because it forces us to consider the evolutionary origins (or side-effects) of the fairness phenomenon. On the face of it, evolution is about genes spreading, not about individuals having balanced life stories. A gazelle that gets eaten by a lion dies in terror and pain—that’s it; nature doesn’t “owe” it any later happiness, and indeed it doesn’t get any. Meanwhile, the lion gets a meal (pleasure) for itself and its cubs. Nature in raw observation often looks cruel and unfair, a point not lost on many biologists.

23.8.2 Conscious experience vs. selection

So is the Law of Fairness at odds with natural selection? Not necessarily, if we frame it correctly. One key point: the Law of Fairness as we propose is mostly about conscious experience, which might not be directly visible to natural selection. Evolution selects for behaviors and capabilities, not happiness per se. It may well be that the balancing of experience is a byproduct—a “spandrel” in evolutionary terms (a feature that is not directly adaptive but comes along with other adaptations). Stephen Jay Gould and Richard Lewontin gave the classic example of the human navel: it exists not because it was selected for its aesthetic or functional value, but as a necessary byproduct of umbilical attachment in the womb. Likewise, the balance of suffering and joy might not itself have been selected for, but could emerge automatically given other things that were selected.

23.8.3 Fairness as an evolutionary byproduct

What other things? Possibly the very mechanisms of neural regulation and homeostasis that keep an organism stable. Evolution definitely favored organisms that maintain internal balance (homeostasis) — stable temperature, stable chemical levels, etc. The brain's reward/pain system evolved to motivate behavior: too much pain and the animal is incapacitated; too much pleasure without end and the animal might not seek food or mates effectively. So, it's adaptive to have checks and balances. For example, the brain releases dopamine for reward, but then has feedback loops to cut it off lest the animal go into a perpetual bliss state and ignore new incentives. Similarly, pain triggers analgesic responses after a certain threshold, otherwise pain could overwhelm and prevent recovery (think of how shock can numb extreme pain during trauma — an evolved response to help you act even when injured).

23.8.4 No extremes allowed

These evolved systems absolutely create a kind of balance. It's not because evolution "cares" about fairness; it cares about function. But in achieving function, it often results in something resembling fairness. One could say: Evolution selected for organisms that don't stay in extreme states, because an animal screaming in pain non-stop is easy prey and can't find food, and an animal in continuous ecstasy might not notice a predator or a drop in temperature. Thus, mechanisms that bring an organism back to a functional baseline (after a high or a low) were selected. Those mechanisms (like opponent processes in the nervous system, hormonal feedback loops, motivation resets) are precisely what manifest to us as "after suffering, relief; after indulgence, satiety/discomfort."

23.8.5 Emergent fairness from stability

In short, evolution may not aim at fairness, but fairness can emerge from evolutionary optimization for stability and efficiency. This aligns with Gould and Lewontin's critique of "adaptationism": not every trait is directly for something; some are byproducts. The feeling of fairness might be a byproduct of multiple regulatory systems each doing their job.

23.8.6 Unselected side-effects

Consider dreams again (from 23.5): one could argue evolution didn't explicitly program "give paraplegics dreams of walking to be fair to them." But evolution did give humans the capacity to dream, likely for other reasons (memory processing, threat rehearsal, etc.). Within that, the human mind incidentally finds a way to simulate wish fulfillment. That might not have been selected; it's just a fortunate side-effect of a complex brain.

23.8.7 Fairness can aid fitness

Another angle: It could be that an organism with a fairness-balanced experience is actually more fit in some contexts. Imagine an animal that endures a painful injury. If no compensatory mechanism kicks in, it might succumb to shock, stop eating, or become depressed (in animals, learned helplessness akin to depression is observed). But if some neurological “reward” comes (maybe via endorphins) to balance that pain, the animal might calm down, tend to its wounds, and survive. That’s directly fitness-enhancing. On the flip side, if an animal experiences a great success (like a big feast), having a subsequent downregulation of pleasure (satiety, maybe a bit of lethargy) prevents it from recklessly seeking more food when it’s already full and perhaps vulnerable. That, too, can be protective. So in many cases, balancing internal states guards against extremes that could be harmful.

23.8.8 Suffering to strength

One might argue, however, that fairness implies excess compensation at times — not just returning to baseline but giving “payback” beyond baseline. Does evolution allow that? Consider phenomena like “post-traumatic growth” again: someone endures hardship and ends up psychologically stronger. At first glance, suffering making you stronger can be adaptive (as Nietzsche wrote, “What does not kill me makes me stronger”; physiologically, stress followed by recovery can build resilience.). Biologists call this *hormesis*: a mild or moderate stressor triggers adaptive repair responses that leave the organism stronger than before (Calabrese & Baldwin, 2002). If that principle holds widely, evolution actually favors organisms that can transform suffering into growth (because they can handle future challenges better). So a certain fairness in outcome (the bad event yields a later advantage) is not anti-Darwinian; it’s a clever survival strategy of the organism.

23.8.9 Fairness in social species

Even altruism and empathy in social species can be seen through a fairness lens: animals (including humans) often help others in distress, which balances out suffering in the group. Evolutionary biologists explain such behavior via kin selection or reciprocal altruism — genes indirectly benefit, or you scratch my back I’ll scratch yours. For example, vampire bats will regurgitate blood to feed a roost-mate who went hungry, an act of reciprocity that ensures no individual starves unfairly (Wilkinson, 1984). Donor bats later receive help in return, illustrating how apparent “fairness” can emerge purely from reciprocal self-interest shaped by natural selection. Those mechanisms can implement fairness at a community level (e.g., in cooperative breeding or mutual

grooming, individuals relieve each other's stress). Here, evolution indirectly "selects for" an outcome that looks fairer for the group because it improves group survival or individual long-term payoffs.

23.8.10 Fairness within a full life

Now, it's true that nature's events can look brutally uneven. Evolution doesn't guarantee anyone a long life; it selects for traits that spread. But if the Law of Fairness is a real constraint on conscious experience, it would apply to whatever span of unified conscious access a creature actually has—however short—imposing its closure condition at the death of mind rather than depending on reproduction or "natural lifespan."

23.8.11 Emergent, not intended

Important distinction: The Law of Fairness is not saying evolution intended to create happiness for every hurt. It is saying that given the systems evolution set up, the emergent result is a balance in experience (much like emergent order in other complex systems). Evolution might not "select against" this emergent result because it usually doesn't harm (and often helps) the organism. And since it doesn't impede survival, it can persist as a neutral or beneficial trait (what in evolutionary theory might be called a neutral drift or a spandrel that became exapted for something, like how maybe dreaming — initially a byproduct — got co-opted for emotional regulation).

23.8.12 Fairness as a byproduct of adaptation

In sum, the criticism that "evolution wouldn't allow a fairness law" underestimates the subtlety of evolution and the distinction between direct adaptations and byproducts. Our view is that the fairness balancing is likely a byproduct of multiple adaptive systems, each ensuring the organism stays within functional bounds, which together give the impression of a moral balance. Nature isn't enforcing fairness out of justice — but in keeping organisms viable, nature incidentally enforces the closure balance the law claims.

23.8.13 Fairness within adaptive limits

To be completely frank, if there were a scenario where strict fairness reduced fitness severely, natural selection could filter that out. In fact, we observe that extreme affective dispositions are generally maladaptive and thus rare. Unchecked mania (relentless euphoria) often leads to reckless behavior and injury, while chronic, unremitting depression hinders survival and reproduction; such persistently extreme mood traits tend to be pruned by natural selection (Keller & Nesse, 2006). Evolution appears to favor

a capacity to return toward an emotional mid-range, avoiding permanent highs or lows. For instance, an organism that compensates every pain with such euphoria that it ignores a mortal threat — that would be maladaptive and likely selected against. And indeed, we don't see that; the compensations are usually proportionate and timed in ways that don't sabotage survival (pain relief kicks in after the immediate danger is dealt with, pleasure down-regulation happens after needs are met, etc.). Evolution has likely tuned the parameters so that fairness effects operate within safe limits.

23.8.14 Conclusion: fairness aligns with survival

Therefore, acknowledging Darwinian logic doesn't invalidate the Law of Fairness; it gives context to it. We interpret fairness in experience as largely aligned with or incidental to survival imperatives. It's an elegant case where what might feel like a moral universe aligning for the individual can be traced to practical biological reason. In other words, the universe's fairness could just be life's homeostasis. For our purposes, that's fine — it still results in the empirical pattern of balanced suffering and happiness. And whether it's a cosmic principle or an evolutionary byproduct, the lived experience is the same: what hurts now is likely to heal, and what delights now will moderate, maintaining an equilibrium that, in hindsight, often feels just.

23.8.15 Where we go next:

Having addressed evolutionary concerns, we arrive at a final critical question: is the Law of Fairness truly testable? In the next section, we grapple with the objection that the law is "unfalsifiable," outlining what evidence could prove it wrong and how we keep our hypothesis scientific.

23.9 “It’s Unfalsifiable”

Objection: “It’s unfalsifiable.” This classic scientific critique targets the very testability of the Law of Fairness. If whenever we point to a case of suffering followed by happiness as evidence, but when faced with a counterexample (someone who dies miserable, say) we might claim the balance happens in some unknown way or in the ‘bigger picture,’ then aren’t we just bending the theory to fit any outcome? A theory that can wiggle out of any disproof is not scientifically valid. Skeptics demand: “What would it take to prove the Law of Fairness wrong? If you can’t answer that clearly, then it’s not a real scientific law, just a comforting idea.”

23.9.1 A restable claim, not hand-waving

This is a crucial challenge because it addresses whether our hypothesis is genuinely empirical or just metaphysical hand-waving. We must be clear: the Law of Fairness is meant to be an empirical claim about conscious experiences. It should be falsifiable in principle. We have to specify what evidence would count against it.

23.9.2 Criterion: lifetime imbalance

One straightforward falsification criterion would be: if we found a large number of individuals whose lifetime experiences show a persistent, unrectified imbalance (either strongly negative or strongly positive) with no sign of compensation, that would challenge the law. We are aware that superficially, there seem to be such cases (tragic lives or perhaps extraordinarily fortunate lives). But we would need to quantify it. For example, if we could measure “total suffering units vs. total happiness units” for a person (perhaps via cumulative self-reports, physiological indices, etc.) and do this for many people, the Law of Fairness predicts these totals should converge to equality at closure (up to measurement noise). If we saw instead a broad distribution with some people ending far in the red (more pain) and others far in the black (more pleasure) and no balancing trends, that would refute or at least strongly undermine the law.

23.9.3 Experiment: induced imbalance

Another potential falsifier: specific experiments could be devised. Suppose we set up conditions to induce an imbalance and see if it self-corrects. For instance, if someone is subjected to extended pain (ethically moderated, e.g., strenuous exercise or cold pressor test) without relief, does their system later generate an extra amount of pleasure (perhaps detectable via mood improvements or endorphin release) beyond baseline recovery? If no compensation above baseline occurs—i.e., they just go back to normal rather than experiencing a rebound high—then strict fairness wasn’t observed in that

scenario. If repeatedly nothing like a balancing overshoot is seen in controlled settings, the strong form of the law might be false.

23.9.4 Counterexample: permanent happiness change

Alternatively, consider extremely happy events: winning the lottery is known to give a spike in happiness that usually regresses to baseline. But what if data showed that some people just stay permanently happier after such events without any drawback or counterweight? If a significant portion of people experience a lasting gain or loss in well-being from life events (past the normal adaptation period), that would mean the scales didn't fully balance; something caused a net change. We'd have to reconcile that or abandon the idea of perfect balance. (There is evidence, for instance, that severe disabilities can cause a lasting drop in happiness, although many people partially adapt. If it's not full adaptation, one could argue fairness wasn't complete. But true falsification would be a gross sustained imbalance.)

23.9.5 Tragic edge cases

Critics might also point to extreme edge cases: say, infant deaths (a baby that lives a short life of mostly discomfort in an ICU and dies—when was its happiness “paid back”?). This is a poignant example. If the law is absolute, one would struggle to see fairness there. Now, one might speculate abstractly about consciousness beyond life or effects on parents (who might cherish their child’s memory, etc.), but that drifts from what we can test. Indeed, such cases put pressure on the idea. If a clear pattern of numerous unbalanced tragedies exists, then unless one introduces untestable realms (like “they’ll be happy in an afterlife”), the law in a strong form is falsified.

23.9.6 Setting a high bar

We intentionally set a high bar by calling it a “Law” and saying it’s a guarantee. That means it should stand up to strong scrutiny. Falsifiability wise, a single apparent counterexample doesn’t kill it, because we might have missed hidden balances (maybe that infant experienced comfort or love in some measure, etc.). But a statistically significant set of counterexamples should. If, for example, we took all people who live and die in extreme poverty and suffering, and found no balancing experiences even in dreams or coping, then the law is in trouble.

23.9.7 Longitudinal tests

In practice, to test fairness, one could utilize longitudinal well-being data. There are studies tracking people’s happiness over decades (like the Grant Study, or national longitudinal surveys). If we examine individuals who report chronically low well-being

and see if any spontaneous improvements or positive experiences intervene over the long term, versus individuals who are chronically high and whether negative experiences bring them down, we can see if trajectories tend to even out. Some research indeed suggests people have a “happiness set-point” they return to, with only small variation. That’s supportive of fairness in broad strokes. But if we found, say, a subset of people who persistently deviated and ended life way below their set-point (without adaptation or rebound), that subset would need explanation.

23.9.8 Defining scope and timescale

A critic might also say we’ve allowed wiggle room by not strictly defining the time scale or scope of balancing. “Maybe in the long run things balance” can always push the ‘long run’ further (even into afterlife or collective consciousness). To keep it scientific, we confine to one lifetime per person. That’s the domain we can measure. Even there, one might push “maybe it balances on the deathbed or psychological resolution at the very end.” That’s somewhat specific: if people near death often experience a sense of peace or acceptance, one could see it as last-minute balancing. If many die in torment, that’s evidence against. So these are observable distributions.

23.9.9 Ultimately falsifiable by data

In short, the law is falsifiable by data on life outcomes and experiences. It stands or falls on whether, when properly measured, the sum total of positive and negative subjective moments equate at closure for each individual stream within the scope of the claim. We have pointed to evidence that such mechanisms exist (opponent processes, hedonic adaptation, etc.). But if future research finds that those mechanisms are limited and many individuals accumulate a surplus of misery or surplus of joy without redress, then we must either modify the theory or reject it.

23.9.10 Statistical vs. absolute law

Because LoF is stated as a strict boundary condition, any verified gross end-of-life imbalance (under preregistered measurement and invariance gates) would falsify it. Our inference is probabilistic because measurements are noisy, but the claim itself is not: a single well-measured violation breaks the law.

23.9.11 Testable predictions

Importantly, the law makes predictions that are testable. For example, it predicts some form of compensatory experience should follow any significant sustained deviation. One could test this in labs (e.g., induce pain, see if pleasure or relief follows beyond normal; induce pleasure, see if discomfort follows like guilt or something). Some of these tests

have essentially been done (the opponent process theory in psychology, drug addiction studies showing pleasure followed by withdrawal pain, etc.). By and large they align with balance. If we find contexts where that doesn't happen—say a new drug that gives pleasure with no comedown or tolerance—then that's an interesting exception (and indeed, addictive drugs are pleasurable up front but nature imposes tolerance/pain later in most cases, which ironically supports fairness). If someone invented a way to short-circuit the brain to have endless pleasure (some worry AI or wireheading could do that), and there was truly no compensatory downside, that would violate the law. So far, even the idea of wireheading (stimulating the brain's pleasure center constantly) meets a physical limit: neurons adapt, or it causes other issues.

23.9.12 No metaphysical escape

In addressing unfalsifiability, we have to avoid moving goalposts. We can't say "Oh, fairness will happen in another dimension or maybe to their soul afterwards," because that's not testable. We've kept this discourse empirical: within a person's conscious lifetime. If fairness doesn't manifest there, the hypothesis fails in its intended domain. We won't escape to metaphysics to save it—that would indeed make it unfalsifiable and thus unscientific.

23.9.13 A scientific hypothesis

Thus, the Law of Fairness stands as a testable scientific hypothesis: we predict that careful, comprehensive accounting of any individual's subjective experience over their lifetime (using as objective measures as possible, like physiological indices of pleasure/pain, behavioral signs, self-reports, etc.) will approach a balance. If such accounting becomes feasible and consistently shows large deviations, then our law would be falsified. On the flip side, every new demonstration of a balancing mechanism (be it neurological, psychological, or social) that aligns with this principle provides further confirmation. We have marshaled many such demonstrations in this book; however, we remain open to the possibility that future evidence could force a revision. This is how any robust scientific claim should stand—strong but not beyond challenge.

23.9.14 Where we go next:

Even a testable law can raise ethical concerns. In the final subsection (23.10), we confront the worry that spreading the idea of an inevitable balance could be dangerous – potentially encouraging complacency or cruelty. Before concluding Part 5, we will address why acknowledging the Law of Fairness does not mean excusing suffering or inaction, which sets the stage for the closing chapter on what it would mean "If Fairness Is Real."

23.10 “It’s Dangerous to Say Suffering Balances”

Objection: “Saying ‘suffering always balances out’ is dangerous.” This is more a moral and practical objection than a logical one. Critics worry that promoting the Law of Fairness could encourage complacency or even cruelty. If people believe that all pain will be compensated by nature, they might become indifferent to others’ suffering (“they’ll get their happiness eventually, so why intervene?”). It could also lead to victim-blaming or trivializing real injustices (“stop complaining about your hardship; it’ll balance out”). In the worst case, a tyrant or abuser might justify inflicting pain by saying the universe will make it right later. The critic thus argues that, regardless of its truth, the idea is ethically perilous and socially irresponsible to spread.

23.10.1 What the law says—and doesn’t

This concern touches on how an idea can be misinterpreted or misused. We should tackle it head-on: Acknowledging a balance in experiences is not the same as condoning suffering or doing nothing about it. The Law of Fairness is descriptive, not prescriptive. It tells us what the theory predicts will happen, not what should happen or what we should do.

23.10.2 Analogy and guardrails

Let’s use an analogy: The knowledge that forests regrow after fires (a natural balance in ecology) doesn’t mean we’d be okay with rampant arson. Yes, the ecosystem might recover in time, but in the interim great harm is done, and human-caused fires can still be evil or preventable. Similarly, even if one believes that someone who is in pain now will eventually experience joy, it doesn’t morally license us to ignore or cause the pain. From a compassionate standpoint, alleviating suffering is still crucial—because suffering hurts now and it’s the humane thing to reduce it. The later compensation doesn’t erase the reality of pain when it’s occurring. Telling someone in agony “don’t worry, you’ll be happy later” is insensitive and unhelpful in the moment (and as the critic notes, potentially insulting or victim-blaming).

23.10.3 Balance can motivate care

In fact, believing in an eventual balance can inspire more empathy and proactive kindness, rather than less. How so? If you think the universe strives toward fairness, you might view yourself as an instrument of that fairness. You might help someone suffering because you feel you’re part of the natural process that delivers their compensatory joy. Many religious or spiritual frameworks that have similar notions (like karma or divine justice) often encourage charitable action: “Be the agent through which balance is restored.” There’s nothing in our law that says “sit back and let the cosmos do it all.” On

the contrary, human actions—since they are part of reality—might be one major way the balance manifests. For instance, one person’s misfortune often rallies support and love from others, which provides the balancing happiness or relief. If those others shrugged and said “oh well, balance will happen on its own” and did nothing, they might actually thwart the mechanism of balance.

23.10.4 No victim-blaming

The cautionary point about victim-blaming is very important. We must clarify: the Law of Fairness does not imply that someone “deserved” their suffering or that it’s somehow okay because they’ll get compensated. This isn’t a moral scorecard where people earn suffering or happiness. It’s a neutral law, like a physical one. If a cruel person inflicts misery on an innocent victim, we are not saying “that victim must have earned this pain and will earn equal joy.” We are saying, independent of any notion of “deserving,” that victim will somewhere, somehow, find experiences that counterbalance, if the law holds. But we still condemn the cruelty and seek justice in human terms.

23.10.5 Comfort without complacency

In fact, knowledge of fairness could be used to comfort without complacency: we can reassure someone that there is hope for future happiness (which can be a very powerful message to prevent despair), while still actively helping them in the present. It’s like telling someone in winter that spring will come (so they don’t give up), but you still give them a coat and firewood for the winter’s night.

23.10.6 Avoid the just-world fallacy

One might recall how the just-world fallacy can make people cruel (“if you suffer, you must have done something wrong”). The Law of Fairness, properly framed, avoids attributing cause or blame; it’s about outcomes. We must take care in communication: stress the guarantee of support or relief, not that suffering is trivial. For example, telling a grieving person “everything happens for a reason, you’ll be okay” can sound dismissive. But saying “I truly believe you will find happiness again, and I’ll be there with you until you do” offers hope and help combined. The difference is empathy and not using fairness as an excuse to distance oneself.

23.10.7 Historical misuse warnings

Consider historical misuses: Perhaps an extreme interpretation of karma made some think the poor or sick were just paying debts, so they withheld charity. That is a perversion of the idea in many religions (most actually encourage compassion, karma or not). Similarly, our secular fairness law should never be an excuse to increase or ignore

suffering on the assumption that “nature will fix it.” First, because part of nature’s fixing might be us doing something; second, because allowing unnecessary suffering is unethical regardless of cosmic outcomes.

23.10.8 Against fatalism

Another dimension is fatalism. If people think everything balances automatically, might they stop striving to improve things? It’s a risk: “Why fight injustice or work hard for happiness if it’s all pre-set to balance?” Our answer: the law doesn’t say when or how balance comes. Human effort could be the very means by which balance is achieved sooner rather than later. Also, fairness doesn’t mean neutrality at every moment; life can have long swings. People still need to act to better their situation and others’. Knowing that ultimately experiences tend to balance can give courage (darkness won’t last forever) but shouldn’t breed inertia (“I can sit and do nothing”). If anything, one might argue it encourages perseverance: if you’re suffering now, keep going because relief is likely ahead—just as you’d keep climbing a hill knowing a downward slope will come eventually.

23.10.9 Safeguards: ethics refrain

A pragmatic safeguard: We should always pair the law’s discussion with an emphasis on compassion and responsibility. Suffering is to be alleviated when possible, not tolerated just because of eventual balance. In our writing and teaching, we clearly differentiate the objective pattern from our ethical stance: The existence of a balancing mechanism in no way diminishes the value of reducing suffering or the duty to be kind. We see it as analogous to pain’s function. Pain has biological value (alerts to harm), but we still use medicine to relieve pain when we can—because unnecessary pain is not morally good even if biologically functional.

23.10.10 Social virtues under LoF

Finally, a positive spin: If widely understood correctly, the Law of Fairness could actually promote a kinder society. People who believe everyone’s joy and suffering are balanced might be less envious of the fortunate and less disdainful of the unfortunate. It could foster a sense of solidarity (“we all go through highs and lows”). It might also reduce despair—people may be less likely to feel hopeless, possibly reducing suicide in extreme pain, if they trust that some countervailing happiness will come. The key is that this trust should not lead them to passivity but to resilience. And those around them, believing the same, would still offer help as instruments of that coming balance.

23.10.11 Where we go next:

Next, in 23.11 we gather the concrete trailheads: measures, datasets, and protocols that let readers audit this chapter's claims for themselves. We point to study families that can separate descriptive balance claims from moral permission, flag known failure modes in interpretation, and show where open data and prereg templates live so that anyone can try to replicate—or refute—what's here.

23.11 Research Notes: Where to Find the Evidence

Subjective experience is measurable. Lauri Nummenmaa and colleagues mapped roughly one hundred feelings into a consistent “feeling space” by combining large-scale surveys with brain imaging. Tor Wager’s team developed a brain-based “neurologic pain signature” using fMRI that tracks pain intensity from neural activity. Together, these lines of work show that the qualitative can be quantified—providing a foundation for empirically tracking suffering and happiness throughout this chapter.

23.11.1 Adaptation and homeostasis

Classic work led by Philip Brickman compared lottery winners with people who became paraplegic and found that, over time, both groups drifted toward ordinary levels of well-being—clear evidence of hedonic adaptation. In the lab, opponent-process dynamics show that ending a painful stimulus evokes a brief surge of relief; with repetition, the primary response diminishes while the after-relief grows. Together, these mechanisms—psychological and physiological—anchor the balance claims used in this chapter.

23.11.2 Just-world vs. fairness

To keep our claim distinct from the “just-world” bias, draw on Melvin Lerner’s program of research on people’s tendency to assume that outcomes are deserved. That bias can encourage victim-blaming. By contrast, the Law of Fairness is a descriptive hypothesis about patterns in experience, not a moral endorsement of what happens. Keeping that distinction clear prevents the idea from being misused as judgment.

23.11.3 Personal identity

Derek Parfit argued that personal identity is not an all-or-nothing essence; what matters are psychological connections across time. That framework supports our stance in §23.4: fairness can operate over evolving or even partitioned selves, so long as there is sufficient continuity in the stream of experience.

23.11.4 Dreams and compensation

Saurat (2011) offers concrete data on paraplegics *dreaming of walking*—strong evidence that dreams supply what waking life lacks. For broader context, see Perogamvros (2023) on dreams regulating emotion by placing threats in safe contexts. These studies (in *Consciousness and Cognition* and *Scientific Reports*, respectively) empirically support Section 23.5’s claim that dreams aren’t random noise but play a role in balancing our emotional ledger.

23.11.5 Predictive brain

To delve into the idea of the brain as a prediction machine, see Hohwy (2013) or Clark (2015). (A quick introduction is the Oxford press release “The brain is a prediction machine,” which encapsulates the idea that our brain constantly anticipates inputs.) This concept, referenced in Section 23.6, helps explain mechanisms but should be contrasted with evidence of real affective balancing (e.g. the dopamine reward-prediction errors in Wolfram Schultz’s work). It shows the brain’s predictive nature, but our law requires more than just prediction – it requires *correction*, as documented in opponent-process and homeostatic regulation research.

23.11.6 Simulations of fairness

While no simulation alone *proves* the law, related computational studies in affective neuroscience or artificial life can be insightful. For instance, computational models of emotion regulation often implicitly create balance to optimize an agent’s performance (see Gross and Jazaieri, 2014, for an overview of emotion-regulation strategies). If you’re interested in the emergence of “fairness,” look at game-theoretic simulations of reciprocal altruism. (Axelrod’s famous 1984 tournament on the evolution of cooperation is a classic example. It’s tangentially relevant: it showed how fairness-like tit-for-tat strategies thrive, hinting that balance has evolutionary advantages.)

23.11.7 Evolutionary perspective

Gould and Lewontin’s famous paper on “The Spandrels of San Marco” is essential to understand how not every trait is an adaptation. We invoked this idea to suggest fairness might be a byproduct of evolution. For concrete examples of evolutionary byproducts, see Nesse and Williams (1994), which discusses how pain and moods can be understood in adaptive terms (useful context for Section 23.8). It’s also worth reading about post-traumatic growth: Tedeschi and Calhoun (2004) document positive changes after trauma, reinforcing that sometimes evolution equips us to rebound stronger.

23.11.8 Falsifiability and empirical studies

For Section 23.9, one should recall Karl Popper’s emphasis on falsifiability in science. More pertinent here are empirical happiness studies. Diener (2006), “Beyond the Hedonic Treadmill,” explores ways happiness can *change* its set-point — a counterpoint to the idea of a guaranteed lifelong balance, and a prompt to consider when and how the law might fail. Longitudinal projects like the Grant Study (Vaillant, 2012), which followed individuals over 70+ years, provide rich data; such archives can be mined to see if real lives truly balance out or not.

23.11.9 Ethical implications

To address Section 23.10's concerns about misuse, one can consult ethics literature. For example, Susan Wolf's critique of "moral saints" or other writings on overly optimistic philosophies give perspective on how not to be naïve or callous about suffering. Works on compassion fatigue and effective altruism also inform how believing "things will balance" intersects with helping behavior. Psychologist Melvin Lerner's research on just-world belief is relevant here too (to avoid subtle victim-blaming in our mindset). In essence, blending the Law of Fairness with humanism requires careful attention, as emphasized by thinkers like Derek Parfit (who balanced hard truths with ethical living) and the Dalai Lama (who often reminds us that regardless of karma, we must act with compassion — a spiritual parallel to our stance).

Each of the above points to rich sources that underpin our chapter's arguments. This body of evidence – spanning neuroscience, psychology, philosophy, and evolutionary biology – can be explored further for a deeper understanding of *why* we assert the Law of Fairness, and also *how* to responsibly apply this understanding in life and science.

23.11.10 Where we go next:

Armed with critiques and clues from these notes, we move to the book's final part. Chapter 24 will imagine *what changes* — in science, care, and daily life — if the Law of Fairness is real. It's also a practical call: whatever the law's truth, there are things we can do right now to foster relief and balance.

Chapter 24 — If Fairness Is Real

If the Law of Fairness is real, even approximately within its known bounds, it would rank among the most consequential ideas in human history. It would not give us a new slogan or moral bromide; it would establish a constraint on lived experience that is measurable, falsifiable, and actionable. Such a discovery would reshape how we practice care, design institutions, educate children, argue in public, grieve and forgive, set policies, and even how we talk about meaning. In short, it would touch practically every aspect of human life.

One way to see the stakes clearly is to run a utopia stress test. Most people imagine utopia as a place where suffering is deleted and the positive spectrum is unbounded. Under LoF, that picture is unstable. If the lifetime ledger must close within the neutrality band, large-scale stability cannot come from deleting cost, it can only come from deciding where the cost lands and in what form. That yields two contrasting design regimes. External Forced Balance is the attempt to keep variance low by narrowing behavior from the outside, using surveillance, penalties, and hard constraints to suppress high-leakage dynamics. Internal Moderation-First is the attempt to keep variance low by teaching people to self-regulate, close loops early, and favor reversible choices without constant enforcement. Both regimes can reduce visible misery, but they fail differently: External Forced Balance tends toward chronic vigilance and a background tax of control, while Internal Moderation-First tends toward fragility unless boundaries protect it from defectors and high-leakage technologies. If LoF is real, utopia is not “maximum pleasure,” it is a narrow equilibrium between volatility and coercion, with repairable trajectories preserved and irreversible traps minimized.

Let’s restate the claim with scientific discipline. Across a single conscious life, the felt sum of experience lands neutral by the end, not due to cosmic intention or anyone “deserving” it, but because multiple guardrails constrain what experiences are possible as time runs out. As horizons H_t shrink and channels C_t remain open, the system naturally shifts toward higher Φ (more reparative, reversible, closure producing) options; these are signatures we test rather than presume. In practical terms, the measured proxy ledger is $\hat{L}(t) = \int_0^t HCl(\tau) d\tau$ with HCl defined as a validated latent composite measure of felt experience (as introduced in Chapter 7). The underlying true ledger $L(T) = \int_0^T F(\tau) d\tau$ is not directly observed, but if LoF holds, the proxy will reflect a trajectory forced toward neutral over a lifetime. Our job is to look for the signatures of that balancing: for example, do we see experience trajectories actually converging to neutrality in the final chapter of life? Do we see that when people’s options narrow, they gravitate to choices that heal or

can be undone? These are the kind of empirical marks a “fairness law” would be expected to leave.

If fairness is real, how would the world change? Quite profoundly, and on many levels. For one, we’d gain a common practical language for care. Religious, spiritual, and scientific communities could talk to each other without surrendering their core beliefs, by focusing on concrete actions: keeping relief channels open, not unnecessarily shortening anyone’s horizon of hope, favoring choices that can be reversed or healed, and helping each life reach a sense of closure. LoF’s secular “grammar” wouldn’t erase anyone’s theology or philosophy; it would bracket those differences and create an immediate common cause around relief, reversibility, and closure for suffering. Justice would subtly shift: courts, schools, and workplaces would put more weight on compensatory and restorative outcomes (without implying moral desert). At equal immediate utility and accountability, solutions that can be repaired or reversed (an apology that genuinely mends harm, a process that can be undone if a mistake is discovered, a consequence that preserves the chance for future redemption) would be preferred over irreversible punishments. Retribution for its own sake would no longer be the silent default. Under the pressure of life’s shrinking horizons, restoration and repair would become the measurable aim of justice.

Markets and education would adjust too. If certain choices predictably strand people on low-Φ (low-feasibility, hard-to-repair) paths, then markets must respond. We’d see sectors like lending, insurance, employment, and design start to price in, or outright forbid, options that create irreparable traps. Conversely, there’d be incentives for “graceful exit” features in contracts and products: generous opt-out periods, draft modes, no-penalty exits and other designs that make reversal or relief easier. Anything deemed “too irreversible” would become a quantifiable risk factor that regulators and markets take seriously. Meanwhile, education would teach horizon literacy. People would learn to recognize when their subjective horizon is short (when life feels urgent or options seem to be narrowing) and to respond by favoring moves that are reversible or restorative, finish something, fix something, sleep on it, reach out for help, rather than rash, irreversible actions. This isn’t moralizing; it’s a practical life skill, an awareness of how timing and reversibility can safeguard one’s life ledger.

We’d also gain a new honesty about suffering. LoF would not excuse harm, nor would it promise any kind of cosmic payback for pain. In fact, it would make that kind of harmful rhetoric scientifically indefensible. What it would allow us to say is different: harm is real; guardrails exist; opening channels tilts things toward repair; and comfort and dignity override data collection. We could talk about suffering with both realism and hope,

acknowledging pain without implying anyone “deserves” it, and focusing on opening paths to relief. End-of-life care, in particular, would take on a new precision. With channels C open and logged, “neutrality” at life’s end becomes a concrete specification, not just a comforting wish. Near terminal time T, we would expect a person’s final days or weeks to show a tightly bounded average experience (for example, a running average whose uncertainty interval lies entirely within ± 0.15 z in standardized HCl units, a short-term slope whose uncertainty interval lies entirely within ± 0.05 z/day, and a variance ratio ≤ 0.80 relative to baseline, as established in 22.5). In practice, families could receive clear, transparent charts of these measures instead of vague assurances, and clinicians would have concrete targets (and accountability) for interventions. End-of-life care would shift toward providing measurable comfort and dignity, treating neutrality of experience as a real outcome to be secured when possible.

Of course, the scientific community wouldn’t stand still. Research practices would shift immediately. The era of single “happiness meters” would give way to an audited multi-indicator model, the HCl, as the default way to measure felt experience. Critically, this HCl model must demonstrate configural-to-metric invariance (and scalar invariance where statistical power allows) across different languages, cultures, and devices. Only with such invariance can cross-group comparisons be trusted; without it, claims would stay within-person. We’d also move from observing horizon effects to manipulating horizons experimentally. Instead of just inferring that someone felt “time is short,” researchers would design studies where participants are deliberately put through Short-Horizon vs. Long-Horizon conditions (with options matched for immediate appeal and including neutral control choices) to see if they reliably shift toward higher- Φ choices when the horizon shrinks. The key measurable interaction is a horizon-by-feasibility effect: do people choose more reversible or reparative options as their time perspective tightens? We hypothesize yes, and we would design studies to test that, using proper statistical checks (for instance, start with Poisson models for count data and switch to a negative binomial model if the dispersion statistic indicates overdispersion (e.g., Pearson $\chi^2/\text{df} > 1.2$)). We’d also never test LoF in isolation: rival models would be our constant companions. Any serious experiment would pit LoF’s predictions against those of alternative theories (say, a pure free-energy minimization model, or a standard reinforcement-learning set-point model). If LoF’s distinctive signatures (like the $\Phi \times \text{horizon}$ interaction or end-of-life “compression”) don’t outperform those rivals at comparable complexity, we don’t declare victory; we revise our theory. Embracing LoF would mean embracing a culture of adversarial collaboration: organized “red-team” challenges to poke holes in the theory would become a normal part of our scientific infrastructure. If a challenge finds a genuine break (say, a situation where LoF fails), the

research protocols get updated, the claim gets revised or narrowed, and the erratum is shared openly. That transparency isn't a concession; it's part of treating LoF as a bona fide discovery rather than anyone's pet idea.

Culturally, none of this would settle debates about the ultimate meaning of life or why the universe permits suffering; LoF doesn't answer those metaphysical questions. What it could do is refocus ethical practice. Religion and philosophy might still grapple with "why," but clergy, counselors, and moral philosophers would hone in on tangible goals: keep channels open, honor closure, prevent irreversible harm when possible. Very different worldviews could share that operational target (relief and closure for individuals) without agreeing on a cosmology. Politics and policy would shift priorities toward interventions that open channels and preserve futures. Resources would flow to things like pain relief, mental health support, communication access, mobility, and direct aid, because these have the highest Φ returns (they make life more repairable). Any law or program would justify itself by measured effects on people's life-ledger trajectory, not by feel-good slogans. In essence, governance would be guided a bit more by "does it measurably keep lives within fairness bounds?" rather than solely by ideology.

Interpersonal ethics would gain new texture. Simple maxims like "be kind" would carry more specific weight. Under LoF, kindness isn't just a virtue, it's an actionable strategy to keep life paths compensable. For instance, if someone's horizon suddenly collapses (they're in crisis or facing a terminal deadline), the ethical move is to default to reversible commitments with them, avoiding trapping them in something they can't undo. Conversely, if you see someone about to do something irreversible while their channels of support seem closed, LoF-informed ethics would urge you to slow them down or open a channel first. These become concrete norms: pace and support others in proportion to how tight their situation is. Collective memory might even adjust. History and storytelling would start highlighting the guardrails in every catastrophe and triumph. Narratives of wars, disasters, or personal tragedies would include chapters on where channels opened just in time, where a reversible choice saved a future, or where someone achieved closure despite the chaos. Similarly, success stories would note how close calls were averted by a timely repair or an open channel. In short, how we remember events would shift: we'd pay as much attention to how balance was preserved or restored (or where it failed) as we do to who won or lost. Societies might become more appreciative of the often invisible acts of repair and relief that keep our ledgers near neutral.

Amid all this, we must keep our claims sober. LoF is a constraint, not a cosmic intention. We intentionally retire any teleological wording. No more "the universe balances things

out” or “everything happens for a reason.” Instead, we stick to the language of constraints: for example, “as horizons shrink, menus tilt toward certain choices.” This keeps us honest: we’re describing a possible natural regularity, not inventing a cosmic purpose. Likewise, we insist on distributions, not anecdotes. We will not “prove” LoF by cherry-picking a few feel-good stories. Lives are not to be reduced to single anecdotes, nor even single metrics. The only acceptable evidence will be statistical distributions that were predicted in advance. We’ll set inclusion criteria, collect data, and see if the distributions of outcomes (choices people make under pressure, end-of-life experience profiles, etc.) match what LoF predicts. Individual stories can inspire hypotheses, but only aggregated, pre-registered patterns can confirm or disconfirm the law. And we’ll be honest about limits. From the start we acknowledge that LoF may have boundary conditions beyond which it fails or falls silent. For example, in situations where channels truly cannot be opened (extreme oppression, irreversible coma, severe untreatable delirium), LoF might not hold, or might not even apply in a meaningful way. Similarly, in trivial short-term scenarios with no emotional stakes, or in cross-species comparisons where a common measurement scale breaks down, we shouldn’t expect the law to operate. Being clear about where we don’t expect fairness to appear is as important as claiming where it does.

Finally, we consider success and failure. What would it look like if LoF is confirmed versus refuted? In the strongest case, we would observe its signatures so reliably that they become decision-grade facts. We’d see replicable horizon-by-feasibility interactions (people reliably favor higher- Φ , more reversible choices as horizons shrink) under careful controls where immediate incentives are tied, and any “neutral” control condition stays flat. We’d see end-of-life experience trajectories consistently meeting the equivalence criteria when channels are open (i.e. near-neutral windows as defined, per 22.5). We’d achieve metric invariance for the HCI across major languages and devices, meaning the measurements truly translate. Rivals would underperform LoF models on out-of-sample predictive metrics. Red-team attacks would probe the theory, but each time the core claims would hold or even get sharper. In short, fairness-in-experience would earn the right to be called a law of nature. Partial success is another possibility: we might confirm some pieces of LoF but not others. Perhaps we reliably see the horizon effect (people do make more reparative choices under short horizons), but the full end-of-life “ledger compression” only shows up under certain conditions, say, only in some illnesses or not at all in cases of acute delirium. We might find that our HCI measurement holds invariance within broad cultural groups but not everywhere. Or maybe some of the rival theories explain certain patterns just as well, leaving LoF only a narrow advantage in specific scenarios (like uniquely predicting dream-rebound effects or very end-game

phenomena). In this scenario, we wouldn't throw LoF out, but we'd downgrade it to a more limited principle, a documented constraint that applies in certain domains rather than a universal law. And yes, failure is an option. LoF could also flat-out fail to hold up. Suppose the best-designed tests show no consistent horizon effect (people don't reliably tilt toward reversible options as time runs out), or suppose even with channels C open we find no neutral convergence at end of life beyond what you'd expect by chance. Or perhaps one of the rival models (like a well-tuned predictive coding model) consistently outpredicts LoF on all key outcomes. If so, then LoF is wrong or at least not broadly true, and that's okay. In fact, knowing that would still be a kind of progress. It would redirect scientists toward whichever dynamics do drive the patterns of suffering and relief in life. Even a refutation of LoF would leave us with sharper questions and better tools for understanding the human condition.

So far, we've considered what LoF might mean for society and science if it's true. But what about you? Section 24.2 turns to "What Ordinary People Can Do." It offers down-to-earth advice: keep the channels of relief open, both for yourself and others. That means seeking help when you're suffering, offering help when you see others in need, and supporting the community practices that make it easier for people to recover (lowering the real-world "cost" of life's counterweights). Law or no law, these actions tilt life toward balance and reduce needless pain. Section 24.3 then issues "A Call for Courage (to Test, Not Believe)." It challenges researchers and readers alike to approach LoF boldly but rigorously. Rather than take this idea on faith or reject it out of hand, we're invited to test it. The chapter suggests concrete steps: preregister your own hypotheses, collaborate with skeptics as well as supporters, and publish what you find. The spirit is genuine inquiry, the goal is to do the work and follow the evidence, not to salute a comforting slogan. Finally, Section 24.4 runs the utopia stress test, and Section 24.5, "Bridge to Synthesis," offers a closing perspective. The bridge emphasizes that whether LoF stands or falls, the goal is to do the work and follow the evidence, not to salute a comforting slogan. We end by looking ahead: the real verdict on fairness will come from future experiments and ongoing compassion, long after this book is finished.

What you'll get from this Chapter:

- The "discovery claim" clarified: We begin by restating the Law of Fairness in clear, testable terms, asserting that every life's ledger must close neutral by the death of mind, and outlining the evidence and exact constraints that make this a scientific claim (not just a philosophical musing).
- Practical guidance if it's true: We discuss what ordinary people can do differently if fairness holds. You'll see why keeping channels open for relief (for yourself and

others), seeking help, offering kindness, and supporting community safety nets are pivotal, small acts that matter regardless of whether the Law ultimately proves true.

- A challenge to test, not just believe: We issue a call for courage to researchers and readers: the invitation is to rigorously test this law. You'll see suggestions for experiments, collaborations (even with skeptics), and open science practices to truly find out if life is fair in this way. The point is to do the work, not to adopt a comforting slogan.
- Utopia as a stress test: We use the utopia thought experiments to distinguish External Forced Balance from Internal Moderation-First, showing how "stability" can relocate costs rather than erase them, and why utopia under LoF is a narrow equilibrium rather than maximal pleasure.
- Bridge to synthesis: We close this chapter by distilling what Chapter 24 uniquely contributes—practical stakes and a discovery-grade frame—and we hand the thread to Chapter 25's final synthesis across disciplines.

Subsections in this Chapter:

- **24.1 The Discovery Claim** - Under LoF the system must close each lifetime ledger at the death of mind; we restate the evidence path and constraints that make this claim empirical.
- **24.2 What Ordinary People Can Do** - Keep channels open for relief, seek help, offer kindness, and support community practices that lower the real-world cost of counterweights; small acts matter regardless of the law's truth.
- **24.3 A Call for Courage (to Test, Not Believe)** - Preregister your own hypotheses, collaborate with rivals, and publish what you find; the invitation is to do the work, not salute a slogan.
- **24.4 Utopia Thought Experiments: External vs. Internal Moderation** - Uses utopia as a stress test to show how stability can be achieved by external enforcement or internal discipline, and why each regime relocates costs rather than deleting them.
- **24.5 Bridge to Synthesis** - Distills Chapter 24's practical stakes and discovery frame, then hands the thread to Chapter 25's final synthesis.

Where we go next:

In this closing part, we begin by clearly stating what exactly is being claimed — and on what terms it would count as *discovered*. That first subsection (24.1) lays out the discovery claim in formal detail, along with the strict criteria and context that guard it.

24.1 The Discovery Claim

What we are claiming — no more, no less.

If the Law of Fairness is real, it's not a comforting slogan but a specific constraint on experience, one that yields distinctive, testable signatures in our data. Our "discovery claim" about LoF is therefore conditional and procedural – it comes with criteria and context, not just a bald assertion.

Claim: Across a single conscious life, the felt sum of experience lands neutral by the *death of mind* — not by anyone's intent or moral desert, but because hidden constraints prune the set of possible life-histories, especially as one's remaining time grows short. As horizons H shrink, the Queue System (QS) prunes the feasible menu toward higher-Φ options (relief, repair, reversibility, flexibility). Open channels simply make this signature legible in data; they are not a precondition for the law's truth. The measured proxy ledger is $\hat{L}(t) = \int_0^t HCl(\tau) d\tau$, where HCl is a validated latent composite of felt experience indicators. (The underlying true ledger $L(T) = \int_0^T F(\tau) d\tau$ is not directly observed.) In plain language: if LoF holds, every life's story must 'balance out' in felt experience by the death of mind; channels matter because they reveal and enable ordinary routes for that signature to appear in measurements, not because the law depends on them.

We only elevate this claim to the status of a discovery if it meets strict confirmatory criteria in controlled tests. In other words, only when all the evidence bars are cleared do we use the word "discovery" without qualifiers. Those criteria are as follows:

24.1.1 Criteria for discovery (confirmatory) horizon × feasibility (menu tilt)

In within-person tasks where we carefully match immediate payoffs and include neutral "both options equal" trials, we expect a specific interaction: as the time horizon (H) shortens, people should disproportionately choose options with higher Φ (greater feasibility for future repair/closure). In technical terms, the interaction coefficient $\beta_{(H:\Phi)}$ should be consistently nonzero in the predicted direction and replicable.

Negative control check: If both available options are equal in Φ (e.g. two equally irreversible or equally reversible choices), choices should not shift — the effect should be flat in those cases.

We will also look for menu tightening behavior: as horizons shrink, a person's "menu" of viable options effectively gets pruned. To quantify this, we examine admissible-set size (a count) while controlling for fatigue and any ledger-position effects, expecting that count to drop. (Methodologically, we'd start with a Poisson model for these counts and

switch to a negative binomial model if the dispersion statistic indicates overdispersion (e.g., Pearson $\chi^2/\text{df} > 1.2$), reporting which model was used.)

End-of-life equivalence (with channels logged). As an individual nears terminal time T, we expect their HCI trajectory (the composite measure of felt experience) to satisfy equivalence criteria *but only if channels are open*. Concretely, under open-channel conditions, the recent window of experience at life's end should have:

- A mean whose uncertainty interval lies entirely within ± 0.15 z of neutral,
- A slope (trend) whose uncertainty interval lies entirely within ± 0.05 z per day (essentially flat), and
- A variance no more than 80% of a comparable mid-life baseline variance (i.e. final days are a bit *less* volatile than the person's earlier baseline).

If those margins are met, we say the ledger has achieved “neutral equivalence.” (We calculate these with proper uncertainty, e.g. using state-space models for smoothing, and pool results across individuals with random effects where appropriate.) If channels are closed during the final window, classify the test as non-confirmatory (indeterminate) and report the channel-open fraction; do not infer failure or any scope limit from such cases.

Measurement invariance (transportability). The HCI measure – our “meter” of felt experience – needs to work across different groups. We require that the measurement model passes from configural to metric invariance across languages, cultures, and data-collection devices (and ideally scalar invariance when sample sizes allow). If we cannot establish at least metric invariance, then any cross-group comparisons or claims of a *general* law are invalid. (In such a case, we would restrict claims to within-person patterns only, which don't require cross-person metric alignment.)

Rival model comparisons (prediction, not rhetoric). We will pre-specify several strong rival explanations for any apparent balancing effect: for example, the predictive coding/free-energy theory (which might argue brains minimize surprise, not ensure balance), a reinforcement learning + homeostasis model (which might say people adapt hedonic set-points without a law), and an adaptation/opponent-process model (which might attribute rebounds to neurochemical opponent processes). For LoF to be considered a discovery, it must not be outperformed by these rivals on out-of-sample predictive metrics for the key endpoints. We'll use measures like WAIC, LOO, or log-loss on held-out data to compare models at matched complexity and regularization. If a rival explains the data better than LoF does, then LoF doesn't get to claim victory just because

it sounds nice; it would mean the “law” might just be an artifact of some known psychological mechanism.

Reproducibility and audit. Finally, an idea isn’t a discovery if only one team in the world can get the result. We require that independent teams, given the same materials and data, can reproduce the key signatures of LoF. This includes running “golden tests” (identical input → identical output) on our analysis pipelines to ensure no hidden hand-tuning. Additionally, we invite adversarial audits: independent analysts should examine our data for issues like privacy or provenance problems (ensuring no subtle data leaks or biases) and test pre-registered falsifier analyses (for example, verifying that some design change *would* have broken the result if the effect were fake). If any of these reproducibility or audit steps fail in a substantive way, we do not declare a confirmed discovery.

Only if all five of the above criteria are met do we allow ourselves to call this a true discovery of a law-like principle. If even one key criterion fails, we either treat the finding as unconfirmed or as only partially confirmed (with a narrower scope).

24.1.2 What we are not claiming

To avoid any misunderstanding, let’s be explicit about several things LoF is *not* saying:

- No teleology. We are *not* claiming there is a purpose or cosmic intent ensuring fairness. LoF is described as a constraint or natural mechanism, not any sort of karmic plan or grand design. There is no “the universe wants this” in our claim.
- No single score defines a life. We do not reduce a human life to one number. The HCI is a latent composite meant for research, not a “happiness score” to label people. A neutral ledger at death doesn’t mean every moment was neutral or that people can be ranked by their life’s sum. It’s about a pattern over time under certain conditions, not a simplification of a person’s worth or happiness into one metric.
- Universality without preconditions. LoF is framed as applying to 100% of unified conscious streams, under all circumstances in which experience occurs. “Channels” (analgesia, sleep, human contact, translation, transport, cash) describe common conduits by which the system expresses balance in practice; they are not preconditions on the law itself. When familiar conduits are blocked or altered (e.g., sedation, isolation), LoF predicts the constraint still holds and may route compensation through other lawful means (changes in salience, valuation, dreams, micro pleasures, and horizon-sensitive adjustments). Cross-person comparisons still require measurement invariance; that is a measurement caveat, not a scope caveat on LoF.

24.1.3 Operational notes (channels as routes, not prerequisites)

Even when all criteria are met, every statement of the LoF claim should come with these boundary conditions clearly noted:

- Channels (C) matter for detection and humane practice. Analgesia, sleep, human contact, translation, transport, and access to basic financial resources (cash) are the best-known conduits through which LoF typically expresses compensation. If channels are not open and logged, any empirical test of LoF is non-confirmatory by design. This is a measurement and ethics constraint; it does not limit the universality of LoF.
- Horizon (H) must be manipulated cleanly to test predictions, not to “enable” the law. LoF’s horizon scaling is a signature we can look for; clean manipulations simply make that signature legible.
- Ledger claims use HCI for accumulation. Our proxy ledger $\tilde{L}(t) = \int_0^t HCl(\tau) d\tau$ is computed from the HCI composite score integrated over time. We always remind ourselves that the underlying true ledger $L(T)$ is unobserved. So whenever we talk about a life’s ledger balancing, we mean the proxy via HCI – unless we eventually discover a way to measure the true $F(t)$ directly. It’s a subtle but important point: we’re currently working with an estimate.
- Ethics override data. At every step, comfort and dignity override data collection. If opening a channel for someone (giving pain relief, letting them sleep, reuniting them with family) conflicts with collecting more data or keeping them in a study condition, we open the channel and abandon the data point. No result is worth violating this principle. (This ethic is built into the law itself: you cannot claim to test a Law of Fairness by doing something patently unfair or inhumane in the process.)

24.1.4 Decision tree for the reader (how to classify a result)

To make all this more concrete, here’s a simple decision tree for interpreting findings related to LoF:

- All five of the above criteria passed? \Rightarrow Declare discovery. (LoF stands as a law under the tested conditions.)
- A key test was non-confirmatory due to closed channels? \Rightarrow Indeterminate (non- confirmatory due to channel closure). Document the channel-open fraction; do not infer failure or scope limits from such cases.

- A rival model clearly wins on OOS prediction, or invariance fails across sites? ⇒ No discovery. (We downgrade the claim or withhold any discovery label. We likely learned something, but LoF doesn't get to be "the law" in that case — at least not yet. We publish the null or the rival's win, and regroup.)
- Any serious privacy/provenance breach that affects conclusions? ⇒ Retract and repair. (We pause any big claims, fix the issue, and issue a correction or retraction as needed. The integrity of the data is part of the discovery; without it, everything else is suspect.)

This outcome-framing ensures that we don't jump to declare a new law of nature unless and until the evidence genuinely warrants it – and that we know exactly what to do if it doesn't.

24.1.5 Public artifacts (what makes the claim auditable)

For transparency, a number of public artifacts will accompany this research so that anyone can audit or replicate the work:

- Preregistration package – A detailed preregistration will include all hypotheses, analysis plans, equivalence margins for tests, criteria for excluding data, a "multiverse" grid of potential analytical variations, the roster of rival models and what would count as them winning, etc., plus a log for any deviations. Essentially, all our cards on the table before we see outcomes.
- Computational containers and seeds – We will release the exact computational environment (containers, code, and random number seeds) that produced our results. Along with that, "golden tests" (specific inputs with known outputs) ensure that anyone running the container can reproduce the same figures and numbers we did, bit-for-bit.
- Tiered data with privacy protections – Data releases will be tiered (perhaps public summary stats, restricted individual-level data under certain agreements, etc.) and come with documentation of how personal identifiers were removed or obfuscated. We'll use techniques like hashing and differential privacy where appropriate, and we'll document data provenance (for example, a triple-hash of code, container, and data snapshot to verify integrity). In short, others should know exactly what data we used and be able to trust that it wasn't tampered with.
- Red-team bounties and errata policy – As mentioned, we'll maintain a public system for others to challenge the findings. If a valid challenge (with data) is submitted, we have a policy to acknowledge it, pay any promised bounty,

downgrade the claim if needed, and publish an erratum or update the next version of results. Essentially, we bake humility and self-correction into the discovery process.

24.1.6 Why this merits the word “discovery”

If all these signatures persist under rigorous conditions – surviving equivalence tests, invariance checks, rival model showdowns, and adversarial audits – then we have uncovered something profound: a law-like constraint on how felt experience unfolds within a life. Such a constraint isn’t just academically interesting; it opens up new predictions, new interventions, and new ethical obligations. It tells us concretely to open channels, to privilege reversibility when time is short, to finish repairs and slow the irreversible, and to measure our impacts honestly. These would become organizing principles in care, policy, and everyday design. And notably, we could do all this without asking anyone to adopt a particular worldview or belief system – we’d simply be building around a discovered regularity of conscious life. That is why, should it all hold, LoF would deserve to be called a *discovery* rather than just an idea.

24.1.7 Where we go next:

We now have a formal claim and a yardstick for discovery. Next comes the question: What do we do with it? In 24.2, we shift from lab-bench criteria to everyday life, showing how even without a single new experiment, ordinary people can start “tilting” their lives toward balance (and why that’s wise whether or not the law is true).

24.2 What Ordinary People Can Do

If the Law of Fairness is approximately true, then when time feels short (when your personal horizon H shrinks) you tend to make better choices by favoring high- Φ options – things that are reparative, reversible, or that close emotional loops. And you do better by opening channels C (making sure you have pain relief, rest, human contact, communication, mobility, and basic resources). The beautiful part is: you don't need a lab or a scientist to help you do these things. You can start using these principles in your daily life right now.

Below is a practical playbook you can use today. These are simple, humane actions and habits informed by LoF's insights. Nothing here is medical advice, and these tips never justify delaying proper care when it's needed — remember, comfort and dignity override any “measurement.” But they're small ways to tilt your own experience (and your community's) toward fairness and relief, starting immediately.

24.2.1 Five moves you can make today (a 10-minute edition)

- Open one channel (C). Think of one “channel” of support or relief you could open *right now*. Pick the easiest: for example, *sleep* – set a gentle alarm to start winding down tonight, or take a short nap if you're exhausted; *contact* – text or call someone you trust just to say hello; *translation* – if language or understanding is a barrier for something you're facing, enable subtitles, use a translate app, or ask someone to interpret; *transport* – make sure you have a ride or add a bit of credit to your transit card so you're not stuck; or *cash access* – set aside a small amount of money (even a few dollars) in a safe place for emergencies. The principle is: open channels first, decide other things later. Many tough situations get drastically easier once a channel for help is open, so do that early.
- Default to reversibility (raise Φ). Before you hit “Send” on that message, or “Buy” on that item, or “Submit” on a decision – pause and ask, “Is there an undo?” If the action is irreversible or hard to undo, see if you can create a reversible step instead. For example, save a draft and set a reminder for an hour later instead of firing off an email in anger. Choose a product with a return policy. Enable an “undo send” delay in your messaging app. If you're about to do something permanent (like a drastic haircut, a big commitment, or even something like surgery), slow down on purpose. Give yourself a day or a week if possible. Making reversibility your default means you'll rarely put yourself on a one-way path without good reason – which keeps your future options more open and balanced.

- Finish one small repair. Identify one little unresolved thing in your life and close it out today. It could be replying to an email you've been avoiding, paying a small lingering bill, cleaning up a spill, or apologizing to someone for a minor mistake. These are “loop closures” that you’ve been carrying mentally or emotionally. By completing even a tiny repair, you reclaim a bit of Φ – you free up future capacity because that task won’t haunt you. High- Φ completions (even small ones) prevent a pile-up of unfinished business that could weigh down your ledger. Think of it as fixing a tiny leak now so you don’t deal with a flood later.
- Do a horizon check. Right now, pause and ask yourself: “What timeline am I really under?” Are you feeling pressure as if you have only minutes to decide something when in fact you have days? Or conversely, are you assuming you have endless time for something that *might* actually be urgent? Name the horizon you’re operating on (minutes, hours, days, months). If you realize it’s short (e.g. a deadline today, or you’re just exhausted and everything feels urgent), choose the option that is most reparative or most easily reversed, even if other options seem equally tempting in the moment. For example: it’s late and you have two leisure choices – one is to answer stressful emails (irreversible time spent, possibly upsetting) and another is to read a book or get some sleep (reparative). Horizon check: your day’s almost done, energy low – better to choose the restorative option. This simple habit can prevent a lot of self-inflicted pain.
- Use kinder language (with yourself and others). Pay attention to your internal narrative and the way you explain situations. Replace any phrasing of self-blame or moral failure with the language of constraints and choices. For instance, instead of “I messed up because I’m weak,” try saying, “Looks like the menu was tight today; next time I’ll try a more reversible move first.” Or instead of saying to someone “You’re being irresponsible,” try “Maybe your horizon is really short right now – what can we do that gives you a way out later?” This shift isn’t about shirking responsibility; it’s about accurately seeing why we sometimes make suboptimal choices (often because our context is constraining us). Using the LoF mindset in language – *menus, horizons, channels* – encourages problem-solving and compassion rather than judgment.

24.2.2 This week: a micro-routine

- The 3-item closure list (daily). At the end of each day, jot down three small things you could close or finish tomorrow. Aim for a mix: one thing to finish (complete a task or project), one thing to fix (address a problem, apologize, or clean up something), and one thing to rest or restore (get sleep, take a mental break). The

next morning, look at your list and tackle the smallest/easiest one first. This habit trains you to seek closure regularly, so that lots of little things don't accumulate into big things. It's like regularly emptying an inbox of life's minor unfinished business.

- One-tap contact. Identify one person who is a “safe” contact for you – someone you could reach out to when you’re in distress or feeling a short horizon. Set up a quick way to contact them: maybe put their number on your phone’s home screen or enable an emergency-call feature. When your day suddenly tightens (you get bad news, you feel overwhelmed), try this: one tap and send a short message like “Short horizon today – can we talk or can you just stay with me on text for a bit?” This is a pre-arranged signal to your friend or loved one that you’re in a tight spot. Even if they can’t do much, the contact itself opens a channel and reminds you that you’re not alone.
- Sleep and pain basics. Treat relief (physical and mental) as a system variable in your life, not a luxury. In practice: set an alarm not just to wake up, but also one to go to bed (reminding you to wind down). Keep simple over-the-counter pain relief handy if it’s appropriate for you, and take it when needed – don’t “tough out” needless pain that can be eased. If you have prescriptions, make a plan so you never run out. And pre-authorize yourself to ask for help or accommodations when you’re in serious pain or seriously exhausted. In LoF terms, opening channels (C) is not indulgence; it’s about feasibility. A life can’t stay balanced if fundamental needs are neglected. So, give yourself permission to use tools like rest, medication, or help from others to maintain your baseline.
- Reversible defaults at work/school. Where you have influence in your workplace or classroom, encourage reversible-first practices. For instance, start projects with a draft phase or prototype that can be critiqued and changed, rather than plunging straight into final products. For meetings, propose having “reversible calendar holds” – tentative dates that can be moved without fuss. For group decisions, maybe implement a 24-hour “cooling-off” period on major commitments, so people can sleep on it. Or use shared documents instead of verbal-only agreements, so there’s a clear record that can be edited. These small policies make the collective environment more forgiving – they assume that sometimes we *all* need to undo or adjust things, and that’s normal. It reduces the fear of irreversibly messing up, which can ironically make everyone more productive and creative.

- Dream counterweights (optional). This one's a bit experimental: If it feels emotionally safe for you, try keeping a dream journal for one week as a personal exploration. Each morning, write 4–6 sentences about any dream you recall, especially noting themes or emotions. See if, on nights after particularly hard days, your dreams feature things like reconciliation, problem-solving, or safe-haven scenarios. You don't have to analyze them deeply or find "meaning" if you don't want to. Just the act of noticing might have an effect: it could subtly steer you toward closure or highlight that your mind is trying to heal or process something. (Some theories suggest that dreams can sometimes act as a pressure-release or compensatory mechanism; whether or not LoF is true, being aware of this may help you be more attuned to your needs.)

24.2.3 With others (family, friends, teams)

- The "open channels" question. When someone you care about is struggling or melting down, instead of immediately giving advice or asking "Why?", try leading with: "*What can we open?*" Specifically, run through the list: Can we get you some rest (*sleep*)? Do you want company or someone to talk to (*contact*)? Do you need help communicating or understanding something (*translation*)? Do you need a ride or to go somewhere (*transport*)? Do you need a little cash or a purchase to solve this (*cash*)? This question can cut through a lot of panic. Often the person hasn't even thought about these basics. Solving one of them first (like "Okay, I added \$10 to your phone so you can make that call" or "Come crash on my couch tonight so you're not alone") can make the bigger problems more manageable. It's a way of being concretely kind.
- Repair script. When you've had a conflict or hurt someone (or been hurt), use a simple repair script to reopen the channel between you. For example: "Yesterday got really tense. I realize I [state my part in it]. I want to repair my part. Here's what I'm doing differently now – does that help or land okay for you?" This script does a few things LoF would encourage: it acknowledges the "tight horizon" moment ("yesterday felt tight"), it takes ownership of one's role in the harm, and it proposes a concrete repair action. Crucially, it asks if the repair *lands* – giving the other person a say in whether closure is achieved. Make the next step easy for them to accept (keep your gesture high-Φ and low-drama). It's amazing how much emotional whiplash can be avoided by swiftly repairing little tears in relationships.
- Undo culture. Foster a culture in your family or team where changing one's mind or course isn't seen as failure but as wisdom. Concretely: maybe have a 24-hour "no questions asked" policy for pulling out of a big decision. If someone says, "I

need to undo that decision I made yesterday,” everyone respects it. Encourage no-fault reversals on first tries: e.g., if someone tries a new role or task and it’s a disaster, let them revert without shame. And ensure there are clear apology and forgiveness paths – for instance, model apologizing publicly when you err, to show it’s okay. An “undo culture” keeps the system fair by preventing one mistake or one bad day from derailing a whole trajectory. It acknowledges we all sometimes need a rewind.

- Meeting design. If you lead meetings (at work, in the community, etc.), tweak the agenda to build in closure and next steps explicitly. For example, start each meeting with 5 minutes reviewing open loops from the last meeting (“Last time we decided X, did we follow through? Any loose ends?”) – this reinforces finishing what you started. End meetings by having each person (or the group as a whole) name one specific closure they will complete before the next meeting. It could be as simple as “I will send that email by Friday” or “We will finalize the budget draft.” Writing these down or saying them out loud means that at the next check-in, there’s a natural accountability to report closure. This practice turns the abstract idea of “finish repairs” into a routine habit for the group.

24.2.4 Digital life, safer by default

- Install friction when H is short. Our devices often tempt us into impulsive, irreversible acts (sending a message in anger, deleting a file, making an unwise purchase) – especially when we’re in a heated or hurried state. Use tech solutions to slow yourself down when needed. For instance, install an email plugin that adds a 10-second delay before a send (with an “undo” option). Turn on recycle bins or trash recovery for your cloud files so nothing actually vanishes immediately. Enable two-step confirmations for purchases or high-stakes posts. These little bits of *friction* are lifesavers for when your horizon shrinks to seconds and you might do something you regret. They effectively lengthen your decision horizon artificially, giving you a chance to catch up with yourself.
- Promote high-Φ choices. Curate your digital environment to make the reparable and reversible actions easier than the one-way actions. For example, pin or bookmark “repair” shortcuts: perhaps a link to your project’s issue tracker (so you’re more likely to log a problem than ignore it), a shortcut to a product return form, or even a template for an apology email. Conversely, de-emphasize or hide one-click irreversible triggers: maybe remove Amazon’s 1-click buy from your toolbar, or add a rule in your social media that saves your post as a draft and requires a second click to actually publish it. By nudging yourself toward actions

that have an “undo,” you maintain more control over your trajectory. It’s like designing your personal UI for fairness.

- Data dignity. In the digital age, one way we can cause unfair harm is by what we record and share about others. An LoF perspective reminds us that privacy is a form of keeping channels open (people won’t seek help if they fear embarrassment) and avoiding irreversible harm (leaks or permanent records can trap someone). So, adopt a practice of minimal and respectful data: avoid storing sensitive or personal information about others in plain text or insecure ways. If you must keep notes (say, as a manager or friend helping someone), keep them protected and only as detailed as absolutely necessary. And never keep “dirt” on someone; if you wouldn’t want it read aloud, think twice about recording it. In simpler terms: treat others’ stories the way you’d want yours treated – with an eye toward their *future* freedom, not just current utility.

24.2.5 When life is acute

- Care first. This is the prime directive. If you or someone you know is at risk of serious harm (self-harm, suicide, harming others), seek immediate help from qualified professionals or emergency services. The Law of Fairness, in those moments, should not even be on your mind – the only law is *safety*. Call emergency numbers, crisis lines, involve doctors or authorities as needed. Fairness or balance is a long-game idea and becomes irrelevant in a true crisis; staying alive and safe is the goal. Once the situation is secure, other principles can come back into play.
- Name short horizons. In an acute situation (a panic attack, a sudden wave of grief, a confrontation), try saying out loud: “Short horizon – high-Φ only.” This can be to yourself or to someone with you. It’s a quick code that means: “Right now, I should only do things that are immediately stabilizing or easily undone.” It might translate to: stop arguing and take a break (because saying more now could cause irreparable damage), or take the prescribed medicine rather than toughing it out, or call a friend rather than making a drastic decision in isolation. By naming the horizon as short, you remind yourself that now is *not* the time for irreversible moves or big gambles. It’s time for safety, comfort, and temporary relief actions until you’re in a better state to decide anything long-term.
- At the bedside (for caregivers). If you are caring for someone in a hospital, hospice, or severe illness situation, use LoF principles as advocacy tools. Continuously ask the question: “Which channel can we open now?” Does the

patient need better pain management (analgesia)? Do they need their loved ones contacted or present (contact)? Do they need translation services to understand what's happening (communication)? Is there an issue with transport or mobility you can solve (transport)? By focusing on these channels, you ensure that the person's experience has the best chance of improving or at least not deteriorating. Also, advocate for comfort measures early – don't wait for someone to be in agony to request pain relief, for example. And finally, if you're logging or journaling on their behalf, never record anything personal that could embarrass them later if revealed. Keep their dignity intact; someday they may recover and read those notes.

24.2.6 A Tiny personal test (optional, 7-day experiment)

You can try a personal mini-experiment with LoF principles without sharing anything publicly. Here's one format for a 7-day self-study:

- Each evening: Write one line rating your overall strain or stress for the day (for example, 0–100, where 100 is the worst stress you can imagine), and note one closure or repair you attempted that day (even if it's very small). This is to see if there's any pattern between effort at closure and perceived stress.
- Each morning: Write 1–2 lines about any dream you recall, focusing on the tone (was it anxious, peaceful, sad, hopeful?) and any major themes (did it involve healing, fixing something, escaping, being helped, etc.). Only do this if it feels safe and not upsetting to you. You're collecting anecdotal data on whether tough days tend to be followed by certain kinds of dreams (LoF would predict some compensatory themes might appear).
- Mid-week and end-week: On two occasions (say Wednesday and Sunday), set up a little choice for yourself when you're feeling pressured. Present yourself with two options that are equally tempting (e.g. two different leisure activities, or two tasks of equal importance), but make one a clearly high-Φ (reversible or repairing) option and the other a more indulgent or one-way option. When you feel your stress/horizon is short, see which option feels easier or more appealing. Take note of it. (This is a micro-test of the “menu tilt” idea in your own experience.)

Throughout this, remember: you're not trying to prove a law on your own – you're just building literacy in how horizons, channels, and feasibility feel in your life. Maybe you'll notice nothing conclusive, or maybe you'll catch one moment where you think “Wow, I really do reach for a reversible choice when I'm in panic mode.” Either way, you're practicing the core ethos: be curious, be kind to yourself, and observe without judgment.

24.2.7 What to teach kids (decision literacy, not dogma)

Even young people can grasp mini-lessons from LoF that will serve them well, without ever mentioning theory or “laws.” Some examples:

- Name horizons. When you’re with a child and a situation is rushed, say explicitly: “We have 10 minutes. What’s a step we can take right now that we can undo if we need to?” This teaches them to recognize time pressure and to prioritize reversible, low-cost actions first. It could be as simple as, “We only have a few minutes, so let’s sketch your ideas in pencil instead of pen, okay?” It instills the idea that a short time frame means you choose differently.
- Repair early. Model and encourage quick apologies and make-good plans. If two kids fight, guide them to apologize and do one nice thing for each other soon after – rather than letting bad feelings fester. If a child breaks something, involve them in fixing it or replacing it (at a level they can handle). The concept to impart: mistakes and harms are normal, but closure should follow soon. It’s not about punishment; it’s about proactively balancing the little ledgers before they become big ledgers.
- Sleep is a tool. Teach kids that being tired or in pain makes everything harder – not as an excuse, but as a reason to take care of those needs. For instance, if homework is a struggle at night, it’s okay to pause and sleep and try in the morning with fresh eyes. Frame rest as part of solving problems: “Our brains work better after a break.” This helps them not see sleep as “doing nothing” or a reward for finishing, but as a legitimate step when they’re stuck or upset. Essentially, they learn that taking care of their body (opening those channels of rest and comfort) is part of finding answers, not a distraction from it.
- (*If they’re a bit older*) Introduce the ledger idea gently. You might explain, in simple terms, that everyone has good and bad days, and we try to help each other so that over time, things feel *more* fair. For example: “Remember when you had a terrible week and then things got better after? Sometimes life works like a balance. Our job is to help keep that balance by being kind when someone’s down.” Keep it practical and avoid any heavy “cosmic justice” language.

24.2.8 How to help at scale (if you want to do more)

- Fund channels. Look around your community or online for ways to directly support the opening of channels for others. This could mean contributing to a micro-grant fund that helps people with transportation or a night of shelter. It could mean donating old devices or laptops to those who lack access

(communication channel), or stocking a public pantry with basic comfort medications or supplies. Even advocating for free Wi-Fi in certain areas is opening a channel (information/communication). These are tangible actions that increase the odds that people in tough times can rebound. If LoF is real, you're literally tightening the fairness guardrails for those folks; if it's not, you're still doing something obviously good.

- Advocate reversibility. Get involved in causes or policies that give people second chances and built-in “undo” options. For example, support legislation that mandates generous return policies or cooling-off periods for major purchases (so people can change their minds without penalty). Back educational policies that allow grade forgiveness or retaking exams. Endorse criminal justice reforms that emphasize expungement of records for minor offenses after rehabilitation. These are all systemic ways of saying “we all make mistakes or impulsive moves; let’s not ruin a life over it.” It’s pushing the world towards fairness by design.
- Share nulls and lessons. If you happen to try any of the small experiments or community projects suggested by this book (like a classroom horizon test or a dream-journal study group), share what you learned with others – even if the result was “nothing happened.” In science, knowing what *doesn’t* work or what patterns *aren’t* there is just as important as positive findings. You could write a short blog post, social media update, or note to the author about what you tried. The point is to contribute to a culture of knowledge where we’re all mapping this terrain together. Maybe your null result will save someone else time, or your clever twist on a small study will inspire a more formal research project. It’s all part of collectively feeling our way toward truth.

24.2.9 Words to keep handy

Sometimes a quick phrase can reset your perspective in a tough moment. Here are a few simple mottos drawn from LoF’s way of thinking – feel free to jot one on a sticky note or your phone for when you need it:

- “*Menus, not morals.*” (Ask “What options do I have?” rather than “What did I do to deserve this?”)
- “*Short horizon → reversible first.*” (When you’re under the gun, default to the choice you can undo or repair later.)
- “*Open a channel before making the call.*” (Before you make a big decision or take a leap, see if there’s a source of support or relief you can tap into.)

- “*Comfort and dignity override data collection.*” (Never forget this guiding rule, no matter what exciting discovery we pursue.)

24.2.10 A note on symbols (for readers seeing lots of notation)

Throughout this book we've used some shorthand symbols. As a quick reference:

- Horizon (H): Think of this as short ↔ long. A short horizon means time is running out or pressure is high; a long horizon means plenty of time or slack.
- Channels (C): These represent those crucial channels of relief (analgesia for pain, sleep, human contact, translation/communication, transport/mobility, and cash/resources). More “C”s open = better.
- Feasibility score (Φ): A higher Φ means an action is more feasible to repair or reverse, or it brings closure. Low- Φ means an action is rigid, irreversible, or leaves loose ends.
- HCl: Hedonic Composite Index – a single latent score combining multiple measures of how someone feels (positive/negative experiences).
- Ledger proxy ($\tilde{L}(t)$): This is the cumulative sum of HCl over time (the proxy for total experienced good-minus-bad). The formula we gave is $\tilde{L}(t) = \int_0^t HCl(\tau) d\tau$. The underlying true ledger $L(T) = \int_0^T F(\tau) d\tau$ (summing actual momentary felt experience) is not directly measured, which is why we call $\tilde{L}(t)$ a proxy.
- Bottom line: Ordinary life already hands you levers to nudge your experience toward fairness – you can open channels of relief, favor reversible and reparative moves, finish small repairs promptly, and slow down anything irreversible when you're under pressure. You don't have to believe in any grand law to benefit from these habits; they stand on their own as good practices. If LoF turns out to be real, you'll be ahead of the curve in aligning with it – and if it isn't, you've still made your life (and others' lives) kinder and more resilient.

24.2.11 Where we go next:

We've seen how LoF's logic can translate into everyday choices. Finally, we invite you to think bigger. In 24.3, the focus shifts to a call for courage in the search for truth – how we, as a scientific and caring community, should test this idea relentlessly (and ethically) rather than simply believe or dismiss it.

24.3 A Call for Courage (to Test, Not Believe)

- Measure, don't sermonize. Wherever possible, replace vague intuitions or moralizing with actual measurements. This book introduced you to tools such as the HCI (composite index of experience) and the notion of a ledger proxy $\hat{L}(t) = \int_0^t HCl(\tau) dt$; consider using them or something similar. Talk in terms of horizons H , channels C , and feasibility Φ instead of saying "I have a bad feeling" or "things will work out." The point is not to become coldly scientific in everyday life, but to practice honesty about what we know. No "vibes" in place of numbers when evaluating LoF – if we say a situation is improving or a person is suffering less, let's have some data or at least a clear observable indicator to back that up.
- Preregister falsifiers in advance. If you're going to test LoF (or any big hypothesis), define ahead of time what outcomes would count *against* it. This is the essence of courage in science: *dare the theory to fail*. For example, before collecting data, state: "If I don't see a $\Phi \times H$ interaction when immediate utilities are matched, I will take that as evidence against LoF," or "If, even with channels open, end-of-life trajectories don't meet the equivalence criteria, that's a strike against the law." Perhaps, "If my HCI measurement doesn't hold metric invariance across these two populations, I can't claim a general law." Write these down and commit to them. By doing this, you ensure that you won't move goalposts later. If the specified outcomes occur, LoF would be considered falsified or in need of serious revision. This protects you (and the field) from the human tendency to rationalize failures away. It's only a law if it stands up to the *possibility* of being proven false.
- Run rivals head-to-head. Don't just test LoF in isolation – that's not a fair fight. Implement the strongest alternative explanations you can think of, and see which predicts reality better. For instance, a lot of people believe "hedonic adaptation" (the idea that we all return to a happiness baseline) could explain balanced life outcomes *without* needing LoF. Fine – figure out what *adaptation* would predict in your experiment and see if that matches or falls short. Or perhaps "the brain is a prediction machine minimizing surprise" – what does *that* theory predict about end-of-life experiences? Put it in code or equations, generate predictions, and compare them to LoF's predictions. Use proper out-of-sample metrics (WAIC, LOO, etc.) so you're not just overfitting one theory or the other. If a rival wins cleanly, have the guts to say "This other model explained things better." That's courageous science – following truth, not allegiance.
- Protect people while you probe ideas. Courage in testing doesn't mean recklessness with lives. On the contrary, truly brave research puts people's well-

being first at every step. Before starting an experiment or intervention, ensure comfort measures are in place: analgesia available, breaks for rest (sleep), someone to call (contact), translation for any language issues, transport if they need to leave, and easy opt-outs (the channels C!). Make it explicit to participants: “You can skip any step or withdraw at any time with no penalty.” Never deceive participants about risks or make them feel they can’t use a “stop” button. The motto is one we keep repeating: *comfort and dignity override data*. If someone gets distressed, you pause or stop and attend to them, period. Designing studies this way isn’t a weakness – it’s a strength. It ensures that if LoF’s ideas are validated, they’re validated in a way that never betrayed the very values of relief and care that underlie the law.

- Publish nulls and accepted hits against us. Have the courage to publicize failure. If you run a solid test and LoF’s signature doesn’t show up – publish that result (or at least share it openly). If you or someone else mounts a “red team” attack and it succeeds in exposing a flaw, don’t sweep it under the rug. On the contrary, pay the promised bounty, thank them, and let the world know: “Here’s where we found a weakness.” We have an explicit policy: any time an adversarial challenge lands a blow, we downgrade the claim and make that known. This isn’t just about honesty; it’s strategic for discovery. It tells others that this problem is hard and we’re not pretending otherwise. It recruits more minds to address the gaps. It shows that what we care about is the truth, not saving face. In the long run, this makes the work unassailable because all the assailable points have been openly exposed.

24.3.1 The three pledges (pick yours)

To solidify the above into personal action, consider formally taking one of these “courage pledges” that aligns with your role. You could even sign and date it for yourself, or share it with colleagues as a commitment.

- Researcher’s Pledge (for scientists): *I will preregister my hypotheses, falsifiers, and analysis plans (the multiverse of ways I might analyze the data). I will blind what I can in my studies and freeze my code and random seeds before peeking at results. I will always compare the Law of Fairness against named rival models using out-of-sample metrics, and I will report those comparisons honestly – if a rival outperforms, I will say so. I will publish my results whether they are positive, negative, or null, along with data (appropriately de-identified and tiered for privacy) and scripts sufficient for others to reproduce my analyses. I will never trade a participant’s comfort or privacy for data; ethics will guide every step of my research.*

- Clinician's/Caregiver's Pledge: *I will treat relief as an essential variable in any care plan. I will document which channels (C) were opened to provide comfort (for example, noting if analgesia was given, if family visits happened, etc.). I will never use the language of “ledgers” or balance to withhold care from someone in pain – if anything, knowing about LoF makes me more committed to actively facilitating balance through care, not assuming it happens magically. When measurement or monitoring is involved in care, I will ensure it is light, fully consented, and never delays or diminishes immediate comfort. In short, I will apply the Law of Fairness only in ways that increase compassion and never as an excuse to neglect it.*
- Citizen's Pledge (for anyone): *In my own life and community, I will favor high-Φ actions when time feels short or when stakes are high – that means I'll lean toward repairing, making amends, creating options that keep the future open, and offering/asking for help. I will strive to open channels before making hard calls: ensuring basics like rest, communication, and support are in place as I face decisions. I will avoid teleological language when talking about suffering (no “it's all for a reason” or “karma will fix it” in serious situations), and instead focus on practical ways to address it. If new and better evidence or arguments come along that challenge what I think about fairness, I pledge to update my views and even publicly acknowledge the change. I'm not here to believe in a comforting idea – I'm here to learn and act on what seems to be true, with empathy and integrity.*

24.3.2 How to be brave without being reckless

- Small, clean, reversible pilots. When testing a big idea like LoF, start with small-scale experiments that are as clean as possible and pose minimal risk. For example, do a within-person trial where one condition is clearly short-horizon and another is long-horizon, and include a neutral control where neither choice has an advantage. Keep the sample small but well-instrumented, match immediate utilities, and pre-register the one outcome you care about (like the $\beta_c H:\Phi$, from Section 24.1.1). By making it reversible (e.g. each participant's involvement is brief and can be halted at any sign of distress) and clean (clear conditions, clear prediction), you gather reliable evidence without overextending or putting anyone in a bind. You can always scale up after a small pilot works; courage isn't about starting huge, it's about starting *right*.
- Equivalence where rhetoric is loud. Some domains (like end-of-life care) are full of strong beliefs and comforting rhetoric. To be brave here means using equivalence testing. Don't just collect data and say “looks balanced, good enough.” Pre-define what would count as “balanced” beforehand (e.g. the ± 0.15

z mean, ± 0.05 z/day slope, etc., as we've used) and then see if your data confidently fall within that range. If they don't, you must say the theory didn't hold in that instance. And absolutely require that channels were open and logged in any end-of-life study – otherwise a failure doesn't tell you much (it could simply be because of closed channels). The bravery here is in holding your theory to a quantifiable standard in precisely the situations where people usually just make comforting noises. It's saying, "We're not going to just *hope* it was fair; we're going to check, and if it wasn't, we'll report that."

- Invariance before inference. This one is for the data wonks: before you compare groups or make grand statements that assume one scale fits all, test measurement invariance. It takes extra time and it's not "sexy" to talk about, but it's the gating factor for fair scientific inference. If you find your HCI metric isn't metric-invariant across, say, two cultures or two languages, don't force an analysis pooling them together and then spin a narrative. Instead, acknowledge the breakdown and limit your claims to within-person or within-group patterns. Courage here means sometimes *not* publishing a big cross-group comparison because you discovered you shouldn't – it means you'd rather have a narrower true claim than a broad faulty one. It's tempting to gloss over these details; a brave researcher doesn't.
- Privacy by design. Build strong privacy and data-protection into your research from the beginning. That means when you design how data will be collected, plan as if you'll have to release it publicly (even if you won't) – how will you anonymize it? How will you prevent re-identification? Use techniques like k-anonymity checks (ensuring no combination of variables can pinpoint a person), consider adding a dash of noise or using differential privacy for any released dataset. Keep personally identifiable information (PII) out of free-text responses and avoid collecting anything you don't absolutely need. Also, use "golden tests" for your code not just for accuracy but to prevent unintended data leaks (for example, make sure that running the analysis twice yields identical results, so you know it's not pulling in uncontrolled external info). When you plan ahead for privacy, you won't be caught in a situation of "We found something amazing but can't share data so no one fully believes us." Instead, you'll be ready to share safely, which strengthens trust in the results.
- Red-team first. Before you finalize any result, especially one that supports LoF, try to break it yourself or invite a colleague to. This could mean applying a totally different analysis method (does the effect still show up if you use a non-

parametric test? a Bayesian approach? add a crucial covariate?). Or deliberately poke at a known weakness: “If I remove channel X from the data, does the pattern disappear? If yes, maybe that channel was doing all the work.” Better you find that out than a critic. Essentially, red-team your own work before publishing. Or literally call up a friendly skeptic and say “Here’s my finding, where would you attack it?” – then address those points. Courage isn’t just pushing your idea through; it’s also about being the first to interrogate it mercilessly.

24.3.3 Scripts for the hard moments

Even with best practices, there will be tough moments in this journey. Here are some ready-made phrases you can use (or adapt) to respond with integrity and courage when those moments arise:

- When a rival outperforms: If you find that a competing model or explanation predicts the data better than LoF does, you might say something like: “Our pre-registered rival model actually explained the primary outcomes better than LoF in out-of-sample tests. In light of this, we are downgrading our confidence in LoF for these signatures and updating the replication kit to focus on understanding this discrepancy.” In simpler terms: acknowledge it and adjust. This communicates to everyone that you’re not clinging to *being* right – you’re clinging to *finding* what’s right.
- When a participant needs relief now: Suppose you’re running a study and a participant is struggling or upset. A good response: “We’re pausing the procedure right here. Your comfort comes first. Let’s take care of you – we can only learn from data that we’d be proud to have collected.” This signals that you value them over the experiment and that data gathered through suffering is not data you want. It also reassures them (and any observers) that the study has ethical guardrails.
- When a result is null: You did the work, and the effect you hoped to see just isn’t there. It’s easy to feel disappointed, but a courageous framing is: “This outcome narrows the range of possible effect sizes for what we’re looking for and gives us a sharper idea for the next test. A null result is not a failure; it’s a data point on the map of reality. Now we know more than we did before we ran the study.” This helps funders, team members, and maybe your own psyche remember that science isn’t only about positive findings. Each null tells us something (even if it tells us to try a different approach next time).
- When a colleague prefers belief over test: If you encounter someone who just *wants* to believe LoF is true (or false) without evidence – say a senior person who

says “I have a gut feeling, let’s not overthink it” – you can gently respond: “Our standard here isn’t belief or gut feeling; it’s prediction, preregistration, and re-run. Let’s make a specific prediction, lock it in, and test it again. That’s how we’ll convince ourselves and others.” This affirms the process and invites them to participate in it, without directly attacking their viewpoint. It redirects the conversation from belief to method, which is where it needs to be.

24.3.4 A minimal courage checklist (printable reminder)

Here’s a concise checklist you might keep at your desk or lab as a reminder of courageous practices. Before declaring a result or submitting a paper, see if you can check all (or most) of these off:

- Hypotheses and falsifiers preregistered
- Seeds/containers frozen; “golden tests” pass (computational reproducibility)
- Immediate utility matched; negative control conditions flat (no hidden biases in design)
- Measurement invariance tested (configural → metric; scalar if needed)
- Channels (C) explicitly logged; participant comfort prioritized; easy “skip” or exit available at all times
- Rival models implemented; out-of-sample prediction metrics reported
- Results + code + processed data archived with DOIs or permanent links
- Red-team invitation or bounty link included (welcoming critique)
- Plain-language summary provided (no teleological language, no moralizing – just the facts and why they matter)

If you’ve got these covered, you’re not just doing science – you’re doing robust, accountable science that deserves the public’s trust.

24.3.5 Why courage matters

Belief can be inspiring, but courage is what protects us from self-deception. In the context of LoF, courage is what stops us from using a beautiful idea as a comforting blanket or an excuse to look away from pain. It forces us to do the hard, slow work to earn the word “law” – or to let it go if it’s not true. The stakes are high: if the signatures of LoF really hold under pressure, we gain a decision-grade constraint on human experience that could reorder care, design, and public life in tremendously positive ways. And if they

don't hold, we still gain better ways of measuring, safer research practices, and clearer thinking about suffering. Either outcome is a win for humanity – but only if we are honest in getting there.

Call to action. So, if you've come this far, consider this a personal challenge: pick a falsifier and test it. Design a small study or observation, preregister exactly what would prove LoF wrong (or right) in that instance, invite a skeptical colleague to review it, and then run the test. Write up what you found, even if it's just a blog post or a note on an observation. We need brave, careful, and reproducible experiments or insights that either make the promise of LoF keepable or show us definitively that we should stop making that promise. In the end, the truth will take care of itself if we have the courage to truly seek it.

24.3.6 Where we go next:

If the Law of Fairness is real, what would a “good world” actually look like under its constraint? Section 24.4 examines two competing control architectures — External Forced Balance and Internal Moderation-First — and asks how each would manage suffering, freedom, and stability without violating the ledger.

24.4 Utopia Thought Experiments: External vs. Internal Moderation

Utopia is a useful stress test because it forces us to say what we mean by “a good world,” and what we mean by “stable.” Many people imagine utopia as a place where suffering has been deleted and the positive spectrum is unbounded. LoF points in a different direction. It treats suffering not as a removable bug, but as a byproduct of any adaptive system living under constraints: prediction, correction, trade-offs, and finite resources. The question becomes: if a society tries to minimize suffering at scale, what control architecture would make that stable, and what costs would it pay to keep the system from drifting into chaos, cruelty, or stagnation?

LoF suggests two distinct answers. One tries to enforce balance from the outside. The other tries to cultivate balance from the inside. They can look similar on the surface because both can reduce visible misery and dampen volatility. But they do it by different means, and they fail in different ways. The point of these thought experiments is not to endorse a political program, but to clarify an engineering question: where does the cost go when you attempt to suppress extreme outcomes?

24.4.1 The thermodynamic framing: utopia is not maximum pleasure

At the collective level, LoF invites a kind of social thermodynamics. When many Unified Conscious Streams interact, you do not just get individual ledgers; you get a coupled system with feedback loops. Some loops concentrate surplus (status, wealth, leverage) and export deficit (stress, humiliation, deprivation). From the outside, “injustice” is not merely an unfair story; in this analogy, it is a high-entropy configuration: a disordered state where large imbalances persist because the system’s corrective pathways are blocked or redirected.

Such a configuration can exist for a while, but it is not free. It requires continual energy input to maintain. That energy can take many forms: coercion, censorship, propaganda, intimidation, legal asymmetries, monopoly control, or the suppression of retaliation and repair. If that input weakens, the system tends to snap toward correction. Revolutions, crashes, sudden prosecutions, and cascades of reputational collapse are examples of imbalances being liquidated when the apparatus holding them in place can no longer pay the bill. In other words, the “correction” is not a moral event. It is the release of stored instability when the containment layer fails.

On this view, utopia is not the state of maximum pleasure. It is a stable regime that handles the waste heat of life, the unavoidable entropy generated by competition, scarcity, conflict, and error, without letting it concentrate into chronic exploitation or

runaway suffering. The design question is not, “How do we create constant bliss?” It is, “How do we keep harm from compounding, and where does the cost land when we try?”

24.4.2 Model A: Forceful Utopia (external forced balance)

The first architecture aims for stability by restricting behavior from the outside. It is the utopia of hard constraints.

Definition: A regime where stability is maintained by strong institutions that narrow the action space of agents.

Mechanism: Surveillance, strict laws, penalties, and restricted access to high-leakage stimuli that reliably destabilize self-control at population scale. Think of bans on certain drugs, aggressive throttling of gambling loops, tight enforcement against predatory behavior, and a general intolerance for behaviors that generate high variance and downstream chaos. This can include “soft force” as well: pervasive monitoring, algorithmic moderation, credential gates, and default friction against risky choices.

Control-theory intuition: This is a high-gain feedback system. When deviations appear, the system pushes back quickly and strongly. It reduces variance by making many destabilizing moves impossible, and by increasing the cost of defecting from the desired regime.

The thermodynamic trade-off is the point. A Forceful Utopia can reduce acute suffering. It can prevent visible spikes: violent crime, obvious exploitation, and social collapse. But LoF’s lens says the valence cost is not eliminated. It is shifted.

What gets shifted is the character of the suffering:

- Acute suffering is reduced, but chronic suffering increases. The chronic form is a low-grade hum: vigilance, lack of freedom, fear of penalties, the sense of being watched, and the psychological tax of living under tight constraint.
- Balance is achieved by a background tax. The system prevents dramatic negative troughs by imposing a continuous, smaller negative load on everyone, or on everyone who might otherwise defect.
- The signature pathology is the Digital Panopticon: a stability regime where the energetic cost of policing becomes the substitute for the energetic cost of disorder. You can get peace, but it is the peace of constant monitoring and narrowed possibility.

This is why an externally enforced utopia often feels, from the inside, like a polite cage. The streets may be safe and the worst harms may be rare, yet the population is still

paying, moment by moment, in diffuse negative valence. LoF's claim is not that such a regime is "bad." It is that it is not free. The ledger still has to balance. The cost changes shape, and the price is paid in autonomy and ambient strain.

24.4.3 Model B: Willing Utopia (internal moderation-first)

The second architecture aims for stability by changing what people do without forcing them. It is the utopia of internalized regulation.

Definition: A regime where stability is maintained because individuals have internalized regulatory policies.

Mechanism: Self-regulation. Individuals voluntarily plug leaks, close open loops, and accept minor discomforts early to avoid major crashes later. In a moderation-first society, many people do not need to be coerced into restraint because restraint is already part of their policy. The system runs on culture, education, and shared norms that make self-correction ordinary, not exceptional.

Control-theory intuition: This is distributed feedback. Each agent acts as their own regulator. Instead of a single high-gain controller pushing everyone back toward the center, thousands or millions of low-gain controllers keep themselves stable from within. Enforcement still exists, but it is less central and less visible because it is not doing all the work.

The thermodynamic trade-off is different here:

- Benefit: Low monitoring costs. Trust can replace surveillance in many domains. Background vigilance and fear can be lower. Transaction costs drop because fewer interactions require enforcement, dispute escalation, or adversarial proof.
- Cost: High developmental cost. A moderation-first society is expensive to build. It requires heavy investment in education, culture, and self-regulation skills, and it requires ongoing expenditure of effort by individuals. The cost is paid metabolically and psychologically, not just institutionally.
- Vulnerability: Free riders. The internal-regulation equilibrium is thermodynamically fragile. If defectors enter and exploit trust, the system can destabilize quickly, and the response tends to be a reversion toward external enforcement.

This is the central tension. A Willing Utopia can feel closer to what people mean by "heaven on earth," not because it has deleted suffering, but because it has reduced the need for coercion. The price is that the population has to be capable of carrying that burden, and the system has to stay intact against destabilizing inputs that overwhelm

internal moderation. Perhaps validation of the Law of Fairness will help bring about a Willing Utopia on Earth.

24.4.4 Heaven and hell as control regimes, not destinations

LoF reframes “heaven” and “hell” as structural regimes of ledger and control, not as metaphysical places.

A heaven-like regime is characterized by high autonomy, low volatility, and low external enforcement. Stability is achieved through internal alignment. The cost is paid in voluntary discipline: people restrain themselves, close open loops, and avoid high-leakage temptations because they understand the long-run costs and have the capacity to act on that understanding.

A hell-like regime is characterized by low autonomy, high volatility, and high external enforcement. Stability, when it exists, is imposed through coercion. The cost is paid in fear and restriction: people remain within bounds because the alternative is punishment, chaos, or predation.

The important point is the transition dynamics. Societies oscillate. When internal regulation fails, through decadence, memetic drift, the corrosion of norms, or the spread of high-leakage technologies, the system tends to drift toward the hell-like regime to restore stability. That drift can look like an increase in policing, surveillance, and hard constraints. It can also look like institutions tightening after a period of disorder, even when the stated aim is protection.

This is not a moral endorsement of any regime. It is a descriptive claim about control. If a system cannot sustain moderation from within, it will either accept higher variance and higher acute suffering, or it will impose stronger external constraints. Under LoF, the ledger does not stop demanding payment. The only question is where the payment lands and in what form.

Both regimes can be pushed to extremes. Hell on Earth could become an automated police state in which monitoring systems track behavior and rapidly apply penalties or restrictions to dampen variance, balancing outcomes quickly at the price of constant oversight. It could also take the form of a controlled facility or simulation that imposes forced “balancing intervals” on a population. Heaven on Earth, by contrast, could become a perfectly designed culture of moderation so effective that all people willingly self-regulate, follow stable routines, and avoid high-leakage temptations with little external pressure. It could even take the form of a voluntary facility or simulation where participants opt into structured “balancing intervals” by choice rather than compulsion.

Even then, LoF predicts the cost cannot vanish. It can only be made smaller, made earlier, made more voluntary, or made more diffuse.

24.4.5 The Secluded Utopia Hypothesis

The fragility of a Willing Utopia leads to a further prediction: stable moderation-first utopias may need to begin as secluded basins of trust, limited to those who can reliably self-regulate.

The reason is simple. A low-enforcement, high-trust equilibrium is vulnerable to contamination by defectors and by high-leakage memetics. In an open system with unlimited variance, destabilizing loops eventually enter. That does not require malice. It is enough that some behaviors and technologies are unusually good at overwhelming self-regulation, or unusually profitable for those willing to exploit trust.

Two kinds of filtering become necessary:

- Memetic filtering: Blocking or tightly limiting high-leakage ideas, technologies, and stimuli that degrade internal regulation. Hyper-palatable foods, addictive algorithms, and engineered compulsion loops matter here because they can swamp moderation policies faster than culture can train them.
- Agent filtering: Excluding individuals who repeatedly exploit the norms of self-regulation, or who import predatory strategies into a low-enforcement setting.

Evolutionary game theory supports the intuition. Cooperative equilibria often require assortativity. Cooperators must be able to group with cooperators, and defectors must be excluded, or at least prevented from harvesting the benefits of trust without paying the cost. If defection cannot be contained, the cooperative regime collapses or hardens into an externally enforced one.

This yields the problem of globalism in its sharpest form. A global Willing Utopia is thermodynamically unstable under LoF because it cannot exclude entropy. Open systems with unlimited variance tend to drift toward defectors or race-to-the-bottom dynamics. If boundaries are prohibited in principle, then the system must compensate in practice by raising enforcement. That is how moderation-first utopias slide toward forceful ones.

In a low-enforcement system, a single deregulated agent can reintroduce high-frequency pleasure loops or avoidance behaviors that overwhelm internal regulators. One person imports an addictive drug into a high-trust community. The community faces a choice: expel the catalyst and restore seclusion or implement policing and slide toward Forceful Utopia. This is why “heaven on earth,” if it exists, rarely looks like a global empire. It looks

like a protected basin of attraction, maintained by constant boundary work against the entropy of the outside world.

24.4.6 Case study: Universe 25 and the failure of unbounded utopia

John B. Calhoun's Universe 25 experiment is often invoked as a parable about utopia, and it is useful here as a cautionary case study of what happens when external stressors are removed without creating stable internal regulation.

The experiment placed mice in an environment with abundant food and water and no predation. In the naive imagination, this is what utopia looks like: remove scarcity and danger, and the population should stabilize in comfort.

It did not. The population did not settle into durable harmony. It collapsed into what Calhoun described as a behavioral sink: rising aggression, disrupted social roles, neglect of offspring, and increasing withdrawal. One striking group, later dubbed "the Beautiful Ones," disengaged from ordinary social life and spent much of their time grooming, insulated from both conflict and reproduction.

The LoF lesson is not that pleasure is bad or that comfort causes collapse. The lesson is that removing external constraints does not automatically yield internal alignment. A system can have abundant resources and still generate extreme negative valence through social dysfunction. It can lose the patterns that make life coherent: stable roles, trusted bonds, meaningful scarcity, and internal policies that keep appetites and conflicts from running away.

In the utopia thought-experiment frame, Universe 25 illustrates a general point: you cannot get stability for free by deleting obvious sources of suffering. If you do not pay the cost in internal moderation and boundary maintenance, you tend to pay it elsewhere, either through coercion (Forceful Utopia) or through disorder and collapse, a high-variance outcome that is utopia in name only.

So the forced-balance versus moderation-first contrast is not a philosophical preference. It is a design fork:

- Force balance externally, and you reduce spikes at the price of a persistent background tax, with the digital panopticon as the characteristic failure mode.
- Build moderation internally, and you can reduce both spikes and surveillance costs, but only if you pay the developmental cost and protect the system from free riders and high-leakage inputs.

LoF does not promise a world of maximal pleasure. It suggests that stable “utopia” is a narrow target: low volatility, low exploitation, high autonomy, and disciplined boundaries. If that sounds less like a fantasy island and more like a hard-won equilibrium, that is the point.

24.4.7 Where we go next:

The next step is simple but demanding: to keep testing the Law of Fairness with the same courage and care we ask of others. Chapter 24 ends not with answers but with resolve—to measure honestly, act compassionately, and keep the question alive. The chapters ahead turn from method to meaning, tracing what this search reveals about science, spirit, and the human story.

24.5 Bridge to Synthesis

This chapter distills a narrow claim with wide consequences. Narrow, because the Law of Fairness (LoF) is not presented as a grand philosophy of life or a solution to every social ill; it is a testable constraint on how admissible experiences shift as horizons (H) shrink and channels (C) open. Wide, because if those constraints are borne out, they ripple into medicine, design, ethics, grief counseling, public policy, and even the everyday language we use with one another.

All the formality you encountered—the HCl composite, the ledger proxy $\hat{L}(t) = \int_0^t HCl(\tau) d\tau$; the notion of an unobserved true ledger $L(T) = \int_0^T F(\tau) d\tau$; the feasibility score Φ ; the admissible-set notation $A(t; \hat{L}, H, C)$; the stringent end-of-life equivalence criteria with channels open—none of that was mere decoration. It was the minimum discipline required to talk about something as profound and delicate as suffering and closure without drifting into either sentimentality or fatalism. Precise language and math were used not to overshadow human stories, but to keep us honest about them.

For a moment, think back to the people we met at the very start of this journey – the child facing an unfair loss, the rockstar whose life seemed undeservedly charmed, the monk seeking peace. We didn’t follow their stories through the chapters, but they have been present as guiding examples. In light of the Law of Fairness, their lives are not “explained” or trivialized; rather, they remind us of why this question matters. If LoF were true, it would suggest that the child’s pain, the rockstar’s fortune, and the monk’s serenity all fit into a larger balancing act shaped by human care and choices. And if LoF turns out false, those stories still stand as testimonies to the complexity of life that any theory must confront. Either way, their experiences – and our empathy for them – set the standard for what any scientific claim about fairness must honor.

24.5.1 What we know now

- Guardrails are measurable. We’ve learned that it’s possible to operationalize the idea of “guardrails” on experience. We can manipulate horizons H in experiments, log which channels C are open, and observe decision patterns. Preliminary evidence suggests that when we shorten someone’s effective horizon in a controlled way (making time feel short or stakes immediate), their choices can tilt toward higher- Φ options if immediate payoffs are held constant. When we remove that manipulation, the effect may diminish or disappear. This tells us that something about imminent endings changes how people weigh their options – a key piece of the puzzle.

- Composite beats single meters. We also discovered that subjective feelings can be measured more responsibly by using multiple indicators combined into a latent composite (our HCI). Any single measure (“rate your happiness 1–10”) is too noisy and too biased by interpretation to serve as a universal meter. But by calibrating a composite and demanding that it behaves (invariantly) across contexts, we get a much more stable signal. In plain terms: feelings *can* be counted, but not with a one-size-fits-all yardstick – it takes a thoughtful assembly of measures that has to be validated at each step. This composite approach can yield a more stable signal than old one-metric approaches, and that’s knowledge we carry forward regardless of LoF’s ultimate fate.
- End-of-life requires equivalence, not slogans. Perhaps one of the most concrete contributions is a new rigor in talking about end-of-life “peace” or balance. We no longer have to rely on platitudes like “they found peace at the end” without evidence. Instead, we have *equivalence tests*: with channels open and logged, we expect a neutral-like trajectory (e.g. recent mean around $0 \pm 0.15 z$). We know what it would look like if a person’s final days were statistically equivalent to neutral drift, and we know how to measure it. Anything less (or more) is by design considered non-confirmatory. In other words, we’ve set a standard: if someone wants to claim a life ended fairly, they need to show it with data or at least transparent criteria. If they can’t, then that claim should not be made lightly. This is a win for clarity and honesty in fields like palliative care.
- Rivals belong in the room. We have firmly established that any serious discussion of LoF must include the rival theories – not as enemies, but as collaborators in finding truth. Throughout the book, we brought in predictive coding, reinforcement learning, adaptation theory, etc., and laid out how each would account for the same observations. We insisted on out-of-sample metrics (WAIC/LOO/log-loss) to compare models, meaning we don’t just fit the story after the fact – we predict new data and see who’s right. This mindset, we hope, will persist in whatever research follows. Even if LoF falls by the wayside, the practice of always testing the competing explanations should remain. It makes the science stronger and the interpretations sharper.

24.5.2 What we don’t know (yet)

- Scope. We still don’t know exactly *where* the Law of Fairness holds and where it breaks. Are there specific illnesses or conditions that are complete deal-breakers for balance? For instance, does severe dementia void the law because the mind can’t process experiences normally? What about extreme lifelong deprivation or

trauma – does LoF apply there, or is it fundamentally a principle of “normal range” life circumstances? We identified some hints (e.g., LoF signatures might be silent in closed-channel environments like solitary imprisonment), but mapping the precise scope – the boundary of admissible trajectories – remains crucial. In other words, if LoF is approximately true, it might be highly conditional, and we need to chart those conditions.

- Mechanism. Assuming LoF’s patterns are observed, *how* are they coming about? We have hypotheses (Queue Systems in the brain, homeostatic mechanisms, etc.), but we don’t definitively know the biological or computational mechanism that would implement the Φ -weighting as horizons shrink. Is it an evolved neural circuit that kicks in during perceived nearing-of-end? Is it a learned strategy that accumulates over life? What role do specific brain regions (like prefrontal control hubs vs. limbic drives) play in steering someone back toward neutral? We admitted early on that LoF is agnostic to mechanism, but science can’t be for long – understanding the *how* will be key to fully trusting and utilizing the law. It remains an open frontier.
- Equity. A profound unknown is how LoF interacts with social structures and inequalities. If LoF’s observable signatures depend on channels being open and logged, then it implicitly assumes a certain level of access and agency. What happens when society itself withholds channels from certain people – due to poverty, discrimination, or policy? Does LoF still “hold” in some twisted way (perhaps through internal psychological adaptation), or does it simply fail because what we’re really seeing is that fairness of experience is as much about policy as about minds? In essence, we must ask: when we measure what looks like a law of nature, are we sometimes just measuring the consequences of social justice or injustice? This question is both scientific and moral, and we don’t have the full answer yet.

24.5.3 Promises we refuse to make

As we conclude, let’s be clear about a few promises we will not make, despite the intriguing evidence for LoF:

- That “suffering balances.” We will never use the language of LoF to tell someone that their personal suffering will *surely* be compensated by later joy. That would be a cruel overreach. LoF, if true, is a statistical or systemic principle – it is not a guarantee that any given individual’s pain is “for a reason” or will be made okay. We refuse to offer that kind of false comfort. Balance, in our usage, is something

to test for and strive for through care – it's not something you tell a bereaved parent or a person in agony as a way to brush off their pain.

- That neutrality is a moral goal. LoF suggests a tendency toward a neutral ledger, but we are not saying that the goal of life or of our actions should be to force everyone to end up “zeroed out.” Neutrality is a statistical tendency, not a commandment or ideal to enforce. In fact, trying to micromanage everyone’s experience to be perfectly balanced could lead to perverse outcomes (imagine denying someone too much happiness because it doesn’t fit the ledger – that would be absurd and unethical). The moral imperatives remain what they always were: reduce suffering, increase well-being, treat people justly. If LoF holds, it adds a layer of understanding, not a new moral scoreboard to hit.
- That data trump dignity. We reiterate one last time: no amount of interesting data or desire to confirm a theory will ever justify compromising someone’s dignity or comfort. Throughout the book we maintained “do not harm” and “do not delay relief” as absolute rules. We promise to continue that in practice. If someone is dying and balancing their ledger, we *still* give them pain meds even if it might “mess up the data” – because dignity comes first. If a measurement device is causing distress, off it goes, LoF be damned. This isn’t just researcher talk; it’s a principle we hope everyone carries: people are not test subjects for our curiosity unless they freely choose to be, and even then, their humanity sets the terms.

24.5.4 Promises we can make

On a more optimistic note, here are a few promises we feel confident making, grounded in the approach this book has taken:

- To measure cleanly or not at all. We commit that any time we make a serious claim under the banner of LoF, it will be backed by data that met the standards we’ve discussed: invariance tested, experiment pre-registered and blinded where possible, and analyzed with appropriate statistics. If we ever find ourselves in a situation where we can’t meet those standards (say, the phenomenon is too hard to measure without ambiguity), we will hold back on making grand conclusions. In short, we won’t half-measure and then spin a tale; we’ll either measure well or exercise restraint.
- To publish nulls and accepted breaks. We’ve baked into our ethos the idea that failure is informative. So we promise to publish significant null results (especially from confirmatory tests) and to give credit to those who find flaws. If someone “breaks” LoF in an experiment, and it’s a solid finding, that result will earn a DOI

in our project's logs and maybe even a bounty reward, not a defensive rebuttal. We'll incorporate that knowledge, adjust our claims, and thank them. This way, the body of evidence stays honest and cumulative, not biased by only positive findings.

- To widen channels first. Whether or not LoF holds, one practical promise we can make is to always prioritize opening channels of help in any intervention or policy. Before we try a fancy new measurement or a psychological trick, we'll ask: have we provided basic relief (*pain treated, sleep enabled, social support engaged*, etc.)? This principle can guide healthcare, education, workplace design – you name it. It's almost like an oath: first, open the channels. By doing so, we maximize the chance of fairness and healing naturally, and we minimize the risk of doing harm by omission.

24.5.5 Call to action

If you're wondering what comes next in concrete terms, here are a few parting suggestions tailored to different spheres:

- Clinics: Start experimenting with adding a “channel check” in medical assessments. It could be as simple as a checklist: “Did we address pain? Sleep? Does this patient have someone to talk to? Do they understand what’s happening? Can they get where they need to go? Are finances keeping them from care?” Logging these as routine *vital signs* would be revolutionary. Also, for those in end-of-life care, try shifting to equivalence reporting: instead of only saying “patient is 6/10 pain today,” also report if their recent trend is stable or not (within that ± 0.05 z/day slope margin, etc.), with the understanding that if channels are open, a stable neutral trend is the goal. And above all, never let a research or monitoring protocol override relief – that’s one area where you as a clinician might sometimes have to advocate *against* a study design, in favor of the patient.
- Educators: Incorporate horizon and channel literacy into teaching, whether formally or informally. For example, in a health or psychology class, discuss how decision-making can change under stress or short timelines, and encourage students to notice that in their own study habits (cramming vs. pacing). Teach them that seeking help or taking a break (opening a channel) in a crisis isn’t cheating or weakness, but often the optimal move. Even younger students can grasp: “If you’re really upset, maybe pause and do something that makes you feel safe, then decide.” Also, consider partnering with researchers to run simple LoF-related studies in schools – students might love to be part of figuring out

something so relatable, like a “balance diary” project or testing whether a kindness intervention affects mood swings.

- Engineers and policy makers: There is a huge design aspect to this. Make reversibility a default in systems: e.g., ensure that any user action that can have big consequences has an “undo” or a grace period. In urban planning or policy, think of “low-Φ traps” – predatory loans, irreversible contracts, even public housing rules that evict someone for one mistake. Identify and mitigate those; it’s a fairness-in-experience issue. For product managers, measure something like “time-to-recover” for your users: if they make a mistake using your service, how long until they can get back to baseline? That’s a metric LoF would suggest optimizing. On the policy side, subsidize or mandate options that preserve compensability: e.g. require lenders to offer paths to catch up on missed payments rather than immediate default. Treat an overly punitive, no-second-chances system as a policy failure – because it’s essentially pushing people off a cliff with no ladder back up.
- Researchers and skeptics: The next steps are clear: take the replication kit (which we will provide openly) and run with it. Literally, run your own analyses. Push the data through alternative models. Propose and test entirely new falsifiers we didn’t think of. The skeptics among you are in the best position to design sharp tests, because you know what you find unconvincing so far – so shore up those weaknesses in a new study. If you’re a grad student or aspiring scientist, there is a ton here that can spin off into dissertations: each element (horizons, dreams, adaptation limits, cross-cultural ledger studies) could be its own multi-year investigation. We invite you to not take our word for anything – fork the project, improve it, criticize it in print. The highest compliment to this work is to engage with it critically.
- (*And for everyone:*) Keep talking about this, but do it carefully. If the idea of LoF resonated with you, share it, debate it, but always with the caveats: it’s a hypothesis, not an excuse; it’s to be *tested*, not *trusted* outright. And if the idea rubs you wrong, that’s fine too – articulate *why*, design a way to show it’s wrong, and share that. The conversation itself, if grounded in empathy and evidence, is part of moving us toward whatever truth is out there.

24.5.6 A simple credo

To wrap up, here’s a short credo – a set of guiding principles – that we hope encapsulates the spirit of this book and can serve as a compass going forward:

- Constraints, not teleology. Always frame observations as natural constraints or mechanisms, not as the universe having intentions. We study menus and guardrails, not destiny or cosmic justice.
- Care before claims. Whenever faced with a choice, prioritize caring for people over proving a point. If there's ever a tension between helping someone and gathering more data, help the person. No contest.
- Rivals at the table. Insist that any idea worth its salt welcomes competition. A law that matters should be beatable in principle – otherwise, it's unfalsifiable or trivial. Bring in the rivals and let them sit at the table of evidence.
- Everything re-runnable. In science (and maybe in life decisions too), hold yourself to the standard: if someone else started with the same inputs, would they get the same result? That means documenting processes, sharing code, and being consistent. It fights the creeping biases and hidden-hand effects.
- Speak plainly. No hiding behind jargon or rosy rhetoric when it really counts. Talk about suffering, hope, and evidence in words that people understand. Don't dress up tough truths or sell comforting falsehoods. Be clear and humble.

If the signatures of LoF survive preregistration, blinding, invariance testing, rival modeling, and red-team fire, then what we've discovered isn't a new uplifting story to believe in – it's a new constraint to build around. In that case, we will design systems that privilege reversibility when time is short, we will fund channels that keep futures open, and we will speak about suffering with the precision and honesty it deserves. And if the signatures fail to hold up, we will have created something rare: a careful map of where a beautiful idea does *not* hold, and better tools and data for whoever tries to answer these questions next.

Either way, there is work we can do today that is safe, humane, and useful. We can widen channels, finish repairs, choose reversibility, measure honestly, and correct ourselves in public when we're wrong. If fairness – in this exact, constrained sense – is real, these habits will align us with it. If it isn't, these habits are still the right way for us to treat one another. Go forth and seek the truth, and in the meantime, be kind.

In closing, the question of life's fairness has led us on a journey through science and daily practice. We have not proven a law of nature beyond doubt, but we have sketched a new way to think and to care – one that demands evidence for our hopes and compassion in our actions. Whether or not each ledger truly balances in the end, we can commit to opening every channel of help, to favoring actions that heal or can be undone, and to

telling the truth about what we find. In a world full of uncertainty, these commitments stand on their own as guiding lights.

Ultimately, the Law of Fairness is more than a scientific proposition – it’s an invitation to approach the human condition with both rigor and empathy. It challenges us to test our most comforting beliefs and to care for others as if our interventions truly matter. If the law stands, we will have discovered something profound about the arc of every life. If it falls, we will still have gained deeper insight into suffering and relief, and a more honest way of grappling with life’s hardest questions.

The end of this book is not the end of the conversation. It is our hope that you, the reader, will carry forward the spirit of inquiry and compassion that drives the Law of Fairness. The search for fairness – literal or not – has already sparked better science and kinder practices. Going forward, every careful experiment, every open conversation, every act of measured kindness becomes part of the answer. Thank you for journeying with us. Let us continue to seek the truth, and let us continue, above all, to be kind to one another.

24.5.7 Where we go next:

We have reached the threshold of our last chapter. The work ahead is not to add new machinery, but to gather what we have built—concepts, measures, objections, and cautions—and hold them to the light at once. Chapter 25 offers a final synthesis: it restates the discovery claim clearly, engages the toughest counterarguments, traces the deepest resonances with science and tradition, and closes with a humane ethic for how to act whether the Law of Fairness stands or falls. With the details behind us, we now step back to see the whole.

Chapter 25 — Final Synthesis

Throughout this book, we have pursued a provocative question: *Is life literally fair in terms of felt experience?* In Chapter 1, we began with the audacious idea that every person's joys and sorrows must sum to zero by life's end – a Law of Fairness (LoF) governing conscious experience. From that starting point, each part of our journey built a framework to examine this hypothesis from all angles. We clarified early on that LoF is not about moral justice or reward for virtue; it posits an objective constraint on subjective experience – a kind of hidden symmetry in the way pleasure and pain distribute over a lifetime. We explored philosophical foundations, distinguishing LoF from comforting myths. We defined key terms like “channels” (pathways for relief or compensation) and “horizons” (time scales over which balance might be achieved), ensuring the hypothesis was precise and testable rather than mystical.

Next, we delved into psychology and neuroscience, uncovering well-documented mechanisms of emotional adaptation and regulation. We saw how people tend to rebound from adversity and come down from elation, hinting at a natural push toward equilibrium. We also confronted the sobering evidence of life's unfairness – tragedies and triumphs that seem unbalanced – using them to refine the conditions under which LoF might hold or fail. Moving forward, we grappled with measurement: Part IV detailed the creation of a Hedonic Composite Index (HCI) to track an individual's net affective state. This rigorous tool combined self-reports, physiological data, and behavioral cues into a single momentary index (HCI), which is then integrated over time into the observable ledger $\bar{L}(t)$. By establishing reliable metrics, we set the stage for empirically testing the LoF hypothesis.

As the book progressed, we examined alternative explanations and rival theories. We considered whether known processes (like homeostatic hedonic set-points or allostatic adjustments) could account for a tendency toward balance without invoking a strict law. We analyzed edge cases – from the impact of dreams on one's emotional ledger to the role of major life transitions – to see if LoF's predictions held firm. In Chapter 20, we confronted two stark scenarios: *What if the law is true?* and *What if the law is false?* We discussed how each outcome would affect our understanding of mind and society. By Chapter 23, we were tying together insights from brain science (the brain as a prediction machine that could implement balance as a local mechanism under the global constraint) and deep psychological observations (how even nightmares or fantasies could serve as pressure valves for emotional excess).

Chapter 24 then turned toward practical implications. We treated LoF as a guiding principle for action: emphasizing the opening of channels (ensuring people have access

to relief, support, and second chances) and advocating policies of reversibility (designing systems that never trap someone in an irrecoverable pit of suffering). We outlined how one might carry the research forward and apply LoF-informed thinking in healthcare, education, engineering, and daily life. Even without final proof of LoF, those practices – keeping paths to recovery open, caring before concluding – were held up as intrinsically good. Chapter 24 ended by presenting a credo of rigor and compassion: to seek evidence relentlessly while never using the hypothesis as an excuse for complacency. In essence, it prepared us ethically and scientifically for the possibility that life’s fairness could be a real phenomenon or an enlightening mirage.

Now, in this final chapter, we bring all these threads together into a comprehensive synthesis. Chapter 25 is the culmination of the book’s ideas – our most insightful, far-reaching, and rigorous analysis of LoF. Here we will examine the law’s implications through multiple lenses, from deep metaphysical questions to hard scientific analogies, from the innermost workings of the psyche to the broadest social and spiritual reflections. We confront the theory with the perspectives of the world’s best thinkers in philosophy, physics, psychology, theology, and ethics, making sure that no discipline’s doubts or insights are overlooked. By the end of this chapter, we aim to present a truly totalizing view of the Law of Fairness: one that acknowledges its elegant appeal, its daunting challenges, and its place in the grand landscape of human understanding.

What you’ll get from this Chapter:

- Grand synthesis of everything covered so far. This chapter surveys the Law of Fairness through five distinct lenses, offering a multi-faceted view. First, in the *Ontological and Metaphysical* perspective, you’ll consider LoF as a fundamental principle of reality. It compares LoF to ancient ideas like karma or fate but strips away moral judgment. You’ll explore the daring notion that subjective experience might obey a conservation-like law – that every conscious life has an invariant “hedonic value” and ask what that implies for the nature of consciousness. The takeaway here is understanding how LoF could be imagined as a law of experience rather than a moral myth (see Section 25.1).
- Analogies between LoF and known natural laws. The chapter asks whether there is a balance law for the mind akin to conservation of energy or thermodynamics. You’ll dive into systems theory and feedback mechanisms – for example, could emotional thermostats or control systems enforce lifetime balance? – and consider information-theoretic angles (how the brain might encode experiences to avoid “hedonic saturation”). This section helps the reader grasp which physical

or mathematical principles might support or contradict the possibility of a hedonic balance rule in conscious life.

- Highlights that show how humans adapt to good and bad events, showing examples of resilience and opponent-process reactions that tend to return people to baseline. We also confront the hardest cases: traumas that never seem to fully heal, lifelong optimists vs. pessimists, and even the logic of suicide within the LoF framework. The goal is to see whether any known psychological processes could naturally produce the balance LoF predicts, and to identify circumstances (developmental, cognitive, social) that facilitate or impede balancing one's emotional ledger. Readers will take away a nuanced view of the mind: understanding both how we often gravitate back toward equilibrium and where that tendency might break down.
- Deep connections between LoF and religious or ethical concepts – for example, comparing it to karma, divine providence, or the moral balance of kindness and guilt. Using a “blueprint” metaphor, the text imagines life with predetermined allotments of joy and pain and asks how much free will we have within those bounds. It also examines how everyday virtues and vices might reflect on one's ledger – for instance, how guilt contributes negative balance and kindness contributes positive balance in ways that resemble spiritual teachings. Crucially, LoF is framed as a neutral natural law in these discussions: it echoes the idea of balance found in many faiths, yet insists it occurs mechanically, without cosmic judgment or teleology. This section helps readers understand how LoF resonates with, yet fundamentally differs from, the moral narratives of human spiritual traditions.
- Considerations of what LoF implies for communities and culture. The chapter discusses cultural differences in beliefs about fate and fairness, and it warns against misinterpretations that could paralyze responsibility – such as thinking “why help anyone if the universe fixes everything?” or “I shouldn't enjoy happiness because I owe debt”. Importantly, it then outlines a positive vision: if a fairness law truly holds, we would design institutions that proactively keep channels to recovery open for everyone (in healthcare, justice, education, etc.), inspiring compassion and long-term thinking. Ethical messaging is emphasized: LoF should be used to empower care and responsibility, not to excuse indifference. The reader is given concrete examples of how policies or daily practices might change – all with the goal of maximizing everyone's chance to balance their ledger.

As a whole, Chapter 25 gives you a broad, balanced understanding of the Law of Fairness. It ties together science, philosophy, and lived experience to show both the allure and the

limits of the idea. You will walk away with a “totalizing” picture: appreciating the elegance of the fairness hypothesis, recognizing its deep challenges, and seeing how it fits (or fails to fit) within modern knowledge. The final reflections (Section 25.6) also remind the reader that whether or not life is inherently fair, the very act of pursuing this question has value – it advances our understanding and spurs compassion. In sum, you get a comprehensive closing synthesis that acknowledges all viewpoints and reaffirms that seeking truth and alleviating suffering remain joint, noble pursuits.

Subsections in this Chapter:

- **25.1 The Ontological and Metaphysical Perspective** – We consider LoF as a fundamental principle of reality governing conscious experience. This section discusses how LoF compares to age-old ideas like karma or fate, yet departs from them by stripping away moral judgment. We explore the daring notion that subjective experience might obey a law-like symmetry, and what it would mean for consciousness to have its own “conservation law” built into the universe’s fabric.
- **25.2 The Physical and Systems Perspective** – We examine analogies between LoF and physical laws, like conservation of energy or thermodynamic cycles. This section asks: Is LoF akin to a natural law of equilibrium for the mind? We delve into systems theory, considering whether feedback mechanisms (like emotional thermostats or control systems) could enforce lifetime balance. We also address information-theoretic angles – how the brain might encode experiences to avoid “hedonic saturation” – and discuss whether any known physical principles support or contradict the possibility of a hedonic balance rule.
- **25.3 The Psychological Perspective** – We gather evidence and theories from psychology and neuroscience that relate to LoF. This section highlights how humans adapt to good and bad events, the phenomena of resilience, opponent-process reactions, and affective forecasting errors that cause us to underestimate our capacity to return to baseline. We confront the hardest cases: traumas that never seem to heal, lifelong optimists and pessimists, and even the logic of suicide within the LoF framework. Throughout, we consider whether known psychological processes could naturally produce the balance LoF predicts, and we discuss the circumstances (developmental, cognitive, and social) that facilitate or impede the balancing of one’s emotional ledger.
- **25.4 Spiritual and Moral Parallels** – Here we connect LoF to religious and ethical frameworks. We draw deep theological parallels, comparing LoF to concepts like the karmic law of balance, the biblical notion of divine providence, or the comforting role of the Holy Spirit in suffering. We use a “blueprint” metaphor to imagine life’s design with predetermined allotments of joy and pain, and consider how much free will we have within those bounds. This section also examines how everyday virtues and vices might reflect on one’s emotional ledger – for instance,

how guilt from wrongdoing or peace from kindness contribute to balances that religions often describe as the wages of sin or virtue. We ask whether LoF, though a neutral natural law in our framing, echoes the moral narratives found in human spiritual traditions.

- **25.5 Societal and Ethical Implications** – In this section, we imagine what it means for communities, cultures, and policies if LoF were true (or even taken seriously as a hypothesis). We discuss cultural differences in attitudes toward fate and fairness – why some cultures might readily accept a notion of balance while others resist it. We caution against fatalistic misinterpretations that could paralyze personal responsibility (“why help anyone if the universe fixes it anyway?”) or encourage self-sabotage (“I’d better not be too happy”). We then outline how a LoF-informed worldview could positively influence society: inspiring proactive compassion, long-term thinking, and structures that help people recover from setbacks. Concrete examples are given of changes in therapy, healthcare, justice, and education that align with maximizing the chances of balance for every individual. We emphasize the ethical messaging needed to ensure LoF empowers care rather than excuses indifference.
- **25.6 Final Reflections** – The chapter closes with a personal and scholarly assessment of the journey we’ve taken. We reflect on the boldness of the LoF endeavor and whether we are prepared to present it to the world’s sharpest critics. We consider the outcomes: what it means if LoF stands up to testing, and what it means if it doesn’t. Ultimately, we reaffirm the value of the search itself – how aiming to prove or falsify LoF has already advanced science and fostered compassion. This final section speaks to the reader directly, inviting a continued inquiry and emphasizing that whether or not life is *inherently* fair, our choices help make it fairer. It is a philosophical farewell that underscores the book’s central message: seeking truth and alleviating suffering are joint, noble pursuits that give meaning to the question of fairness.

Where we go next:

With this roadmap in mind, we will weave together the insights from all previous chapters, confronting our hypothesis with every tool and insight at our disposal. In subsection 25.1 we return to the very heart of fairness as a law – first by asking if such a law could plausibly exist at the fundamental level of reality, and what that implies for everything else.

25.1 The Ontological and Metaphysical Perspective

The Law of Fairness was born as a metaphysical proposition as much as a scientific hypothesis. It asserts a law-like regularity in the universe of subjective experience: a claim that every conscious life is bound by a strict terminal boundary condition on the integrated ledger of felt experience. This section examines that claim in its broadest, most philosophical form. Is LoF even *conceivable* as a fundamental principle of reality, akin to gravity or thermodynamics, but operating in the realm of the mind? What would it mean for life to be “fair” by design, and can we frame this idea without invoking mysticism or moralizing? Here, we explore LoF in the context of age-old human questions about fate, justice, and the nature of consciousness.

25.1.1 A law of experience

At its core, LoF posits that conscious experience obeys a conservation-like boundary condition – that the total “hedonic value” (pleasure minus pain) accrued over a lifetime is constrained to be zero net at terminal closure. This is a daring ontological claim. It suggests that the universe has an objective constraint on subjective life trajectories, in the sense of a boundary condition on cumulative affect that must be satisfied at the stopping time T. Throughout the book we’ve treated LoF as a hypothesis to be tested, but here we consider its status as a putative law of nature. Unlike familiar physical laws, which govern mass or energy, LoF would govern the intangible currency of happiness and suffering. It implies that beyond all the visible chaos and injustice of life, there is an invisible symmetry operating on each person’s cumulative experiences. In proposing LoF, we are essentially asking: *Is there a hidden order in the universe that ensures every conscious being’s ledger of joy and sorrow comes out even?*

25.1.2 Beyond karma and teleology

The notion that life’s ups and downs might balance out is not entirely new; it resonates with ancient ideas like karma, the Wheel of Fortune, or the adage “what goes up must come down.” Many cultures and religions have hoped for some form of cosmic fairness. However, LoF is distinct in crucial ways. It is strictly amoral and non-teleological – it does not claim that good deeds are rewarded with happiness or evil deeds with suffering (as karma or divine justice would), nor that the universe *intends* any personal growth or punishment. Instead, LoF is posited as a neutral mechanistic balance: each person gets a mix of experiences that, in total, neither privileges nor torments them more than any other life. In other words, everyone’s felt experience sums to the same endpoint, regardless of their virtue or vice. This is a stark departure from the “just-world hypothesis” that our psychology often gravitates towards (the comforting belief that people get what

they deserve). LoF does not say people deserve their joys or sorrows – only that they inevitably have a proportionate share of both by life’s conclusion.

By stripping away moral causation, LoF aligns more with a principle of symmetry than with traditional spiritual doctrines. If one were to draw a parallel, LoF might be seen as a cousin of karma but without the ethical accounting – a kind of *existential karma* where the balance is kept by nature automatically, not by judgment of one’s actions. It also echoes the yin-yang concept from Eastern philosophy: within every happiness there is a seed of future sorrow and vice versa, maintaining an overall harmony. But again, LoF doesn’t invoke a cosmic judge or purposeful force; it posits a non-teleological boundary constraint on cumulative valence. This absence of teleology is intentional – throughout our exploration, we avoided suggesting that “the universe wants” to teach lessons or enforce justice. We consider LoF a constraint condition or design feature, not a moral plan. In this sense, it is closer to a law of physics (impersonal and exceptionless if true) than to a moral law or divine providence.

25.1.3 Consciousness and the fabric of reality

If LoF is true at this fundamental level, it carries profound implications for how we view consciousness in the cosmos. It would imply that conscious experience has an invariant built into its dynamics, much like energy is invariant in physical processes. In physics, Emmy Noether’s theorem links conservation laws to continuous symmetries of the underlying laws. One might whimsically ask: *What symmetry of the universe could yield the conservation of net happiness?* Perhaps a symmetry between pain and pleasure as two sides of the same coin, or between the beginning and end states of a life. These ideas are speculative, but they show how LoF pushes us to consider consciousness as not just an emergent accident, but something woven into reality with its own rules. To entertain LoF seriously, we must imagine that subjective experience has quantitative structure that nature could regulate. This challenges a strictly materialist view that treats consciousness as a by-product with no universal properties. If a law like LoF exists, it suggests that the realm of the subjective might be undergirded by patterns as regular as those in the physical realm.

Of course, this is a controversial stance. Critics would argue that, unlike mass or energy, we have no evidence that “happiness units” or “suffering units” are a conserved quantity. Feelings are not like particles that can be counted or measured in absolute terms across individuals. Admittedly, proposing a conservation law for something as qualitative as experience borders on metaphysics. We are essentially hypothesizing that the total “hedonic charge” of a closed system (a single life) is zero – or in plainer terms, that each life’s total good and bad cancel perfectly. This raises eyebrows because it posits a hidden

order in what many assume to be random or unjust. It hints that consciousness might follow rules we haven't yet catalogued, possibly requiring an expanded scientific framework that bridges physics and psychology.

Yet, by formulating LoF as a hypothesis, we have taken it *out* of the purely metaphysical realm and into the empirical realm. This is crucial: we are not asserting this law as a given truth or dogma; we are proposing it as a testable claim about reality. In doing so, we acknowledge that if the claim fails rigorous tests, it must be discarded like any other falsified theory. LoF does not ask for faith – it asks for scrutiny. In that sense, even if it has a whiff of metaphysical boldness, it stands apart from unfalsifiable beliefs. We have treated it not as “wishful thinking” but as a bold conjecture that demands evidence. By bringing tools of science to a question traditionally reserved for religion or philosophy, we have, in effect, forced this lofty idea down to earth where it can be examined.

25.1.4 A Design or an emergent pattern?

One tantalizing question is whether LoF, if true, implies anything about design or purpose in the universe. Some readers might wonder: if every life is balanced in the end, does that hint at a benevolent design – a kind of built-in mercy ensuring no one suffers without relief? It's an emotional temptation to think so. Indeed, if LoF were proven, many would interpret it through their own frameworks: a religious person might see it as evidence of a compassionate God or a cosmic plan (perhaps the way holy scriptures say “God will not give you more than you can bear”), while a secular thinker might marvel at the ingenuity of evolution or nature in self-correcting our emotional extremes. It is reasonable to want to lean away from invoking an intelligent designer or mystical enforcement. Instead, if the LoF holds, it could be viewed as an emergent property of complex systems – an outcome of many smaller processes that collectively yield an overall balance. That is, rather than the universe *choosing* to be fair, it could be that the structure of our biology and environment naturally produces an equilibration over long timescales (much as many chaotic processes still have statistical regularities).

The line between emergent pattern and “design” can blur in interpretation. If something consistently holds true and benefits conscious beings (like preventing infinite suffering), one might poetically call it a “merciful design” of nature. In our strictly scientific narrative, we refrained from such language, but in this concluding reflection it's worth noting the theological parallel: LoF would in effect be a guarantee that “*no one suffers beyond their measure*” – a concept not far from religious assurances of cosmic justice or divine mercy. The difference is that LoF lacks any moral targeting: it doesn't ensure the wicked suffer or the good prosper; it only ensures everyone gets a comparable mix. To a deity-focused worldview, that might seem an odd sort of fairness – not aligned with merit, just a blanket

equalization. Some theological interpretations might say LoF is evidence of a divine *impartiality* or a testing ground where everyone is given equal total measure of joy and sorrow regardless of their earthly deeds (leaving ultimate justice to an afterlife, perhaps). We won't venture deeply into those interpretations here, but it's fascinating that an idea so mechanistic in our formulation can evoke age-old religious themes in a new light.

25.1.5 Facing the profound skepticism

Proposing a new fundamental law is always met with skepticism, and rightly so. In the grand scheme, LoF is a radical idea because it touches on the Problem of Evil (why is there suffering?) and flips it into a hypothesis of equilibrium. Philosophers might object that we are inadvertently reintroducing a form of theodicy (justification of suffering) by suggesting it's balanced – something many are uneasy about. Let's be clear: LoF does not *justify* suffering; it attempts to *account* for it. It says suffering occurs and is matched by joy, but it doesn't say suffering is "good" or necessary for some cosmic moral reason. This distinction is subtle but important in philosophical discourse. If anything, LoF intensifies the mystery: it asserts a perfect balance yet offers no why beyond the existence of a natural constraint. It is almost a *brute fact* kind of claim – "that's just how conscious experience unfolds, uniformly for all". In philosophy, one would question whether such a claim can be reconciled with the chaotic, contingent nature of life events. Is it not far-fetched to think that regardless of random disease, crime, luck, and personal choices, everyone ends up emotionally equal? It does sound far-fetched! And this is why we have taken pains to frame LoF as an extraordinary hypothesis requiring extraordinary evidence.

In summary, viewing LoF from an ontological perspective underscores both its boldness and its transformative potential. If true, it would hint that the universe has a hidden symmetry for minds – a symmetry that might arise from deep structural reasons. It would bridge subjective experience with the kind of invariants we see in physics, suggesting a new layer of lawfulness in nature. If false, the exercise of considering it still has value: it sharpened our definitions of fairness and forced an interdisciplinary examination of life's dynamics.

25.1.6 Where we go next:

Having sketched what LoF means in theory (and how it contrasts with familiar moral or religious fairness concepts), we now turn to the empirical analogies and scientific models that either support or challenge the plausibility of such a law.

25.2 The Physical and Systems Perspective

To treat LoF as more than a poetic idea, we looked for analogies in the physical sciences and systems theory. Could the balancing of life's joys and sorrows be akin to known principles like energy conservation, equilibrium dynamics, or control systems maintaining stability? In this section, we explore how far the comparison can be stretched and where it breaks down. We consider whether a human life can be modeled as a kind of *closed system* with respect to emotional energy, and if so, what mechanism would enforce the zero-sum outcome. We also investigate engineered systems (like thermostats and feedback controllers) for insight into how a “fairness regulator” might function. Additionally, we incorporate an information-theoretic viewpoint: is there a reason a finite brain would naturally even out extreme experiences to maximize information or prevent overload? While analogies are not proofs, they provide a structured way to think about LoF. By articulating the parallels with physics and systems, we can derive testable predictions (e.g. statistical signatures of balance) and identify potential “forces” or processes that could make the law work – or reasons it might inevitably fail in a messy open system like human life.

25.2.1 Conservation analogy

In physics, a conservation law means that a certain quantity (like energy or momentum) remains constant in an isolated system. At first glance, LoF sounds like it’s proposing a conservation of “hedonic energy” – that the time-integrated balance of happiness minus suffering in a closed system (one person’s life) is constrained to be zero at terminal closure. However, there is a key difference: LoF doesn’t claim the quantity is fixed at every moment (clearly, our happiness levels fluctuate wildly over time); it claims that the integral over the entire life is zero. This is more akin to saying that the net change over a complete cycle is zero. An analogy can be made to a physical process: imagine a spacecraft that leaves Earth, travels through space, and returns – its displacement is zero at the end of the journey, even though it wandered far from home in between. Similarly, LoF says a life may wander through extreme highs and lows, but by the end, the total displacement from the emotional “zero point” is nil – you’re back where you started in a cumulative sense.

We likened this, cautiously and by analogy only, to a closed loop in a conservative field. For example, in a gravitational field, if you start at a given position (or, more generally, a given gravitational potential) and return to it, the net work done by gravity over the trip is zero (because gravity is conservative). By analogy, LoF envisions life’s emotional experiences as if they occur in a conservative “hedonic field” – no matter the path (sequence of events), the net work (cumulative emotion) is zero if you come full circle to

neutrality at death. This was more than a poetic comparison; it led us to think in terms of symmetry and invariants. We even mused about a hypothetical “Noether’s theorem for subjective experience” in Chapter 3, speculating that if there is some symmetry in the equations governing conscious processes, a conservation of net affect could result. But let’s temper the analogy: conscious minds are not closed physical systems in the strict sense. People exchange matter, energy, and information with the environment constantly – so the isolation needed for a strict conservation law doesn’t obviously apply. Each person influences others (we’ll address later whether LoF might have a social component), and random external events (disease, accidents) dump “emotional energy” into a life from outside with no immediate compensation. Therefore, if LoF holds, it’s not because a person is physically isolated, but perhaps because the *psychological processes* within that person act to counterbalance any influx of joy or sorrow over time. In other words, the “system” that might be closed is the mind’s accounting of experience, not the external events themselves.

25.2.2 Thermodynamics and equilibrium

Another perspective is thermodynamic. One might ask if LoF is akin to a statement of equilibrium for emotional states. In thermodynamics, systems tend toward equilibrium (for instance, hot and cold objects in contact will equalize in temperature). At true equilibrium, every forward process is exactly balanced by a reverse process — a condition called *detailed balance* in statistical physics (Tolman, 1938). There is no net energy flow because each microscopic transition is matched by its opposite. By analogy, if LoF imposes an equilibrium on the emotional life, it implies that for every “forward” surge of positive or negative experience, an equivalent “reverse” experience eventually occurs to cancel it out, yielding no net hedonic change by the end of life. Is life similarly tending toward a neutral emotional state as an end condition? LoF’s claim is stronger than ordinary equilibrium: it’s not just saying we all end up emotionally lukewarm – it’s saying the total heat absorbed equals the total heat released, metaphorically speaking. It implies a *perfect compensation* of every deviation. In thermodynamics, a perfect cycle (like an idealized Carnot engine running in reverse and forward) can net zero change in energy if it’s truly cyclic. But real systems have entropy – some energy is always lost as unusable heat, preventing perfect reversibility. If we translate that to life, one could argue there is an “entropy of emotion.” Perhaps every experience can’t be perfectly undone; maybe some suffering leaves scars that no joy fully removes (akin to dissipated heat), or some happiness lifts us in ways no subsequent sorrow fully drags down. If so, real lives would fall short of perfect balance, more like a damped oscillation that might still leave some residual imbalance at the end.

LoF in its ideal form would be like a 100% efficient emotional engine – no net loss or gain, everything accounted. That rings suspiciously idealized. Indeed, one of the scientific challenges we've acknowledged is to check whether data show a narrowing of imbalance over time or whether some people's "emotional entropy" accumulates without bound. When we discuss empirical tests, we'll be looking for signs of convergence to zero vs. divergence. The thermodynamic analogy also extends to the idea of limits: just as the second law of thermodynamics forbids complete conversion of heat to work without losses, maybe there is a principle that forbids complete balancing of emotions in extreme circumstances. However, it's worth noting that even thermodynamic laws are statistical in nature. Modern fluctuation theorems show that, in small systems over short times, transient negative fluctuations in entropy production can occur (Wang, 2002). LoF, however, is not framed as a statistical tendency. As stated, any verified gross end-of-life imbalance under preregistered measurement and invariance gates would falsify it. If LoF turns out false, it might be because some processes (e.g. neurobiological damage from trauma) introduce irreversibilities – like irreversible entropy – that prevent a full return to neutral. We keep this caution in mind even as we explore the elegance of the balancing idea.

One can extend the thermodynamic perspective by noting that life exploits kinetic constraints. Pross and colleagues describe *dynamic kinetic stability* (DKS) as the persistence of a replicating system in a cyclical, energy-driven steady state. In DKS, a replicator population reaches a steady size only when reproduction exactly balances decay, so that faster-reproducing entities dominate in a blind, selection-like process. This kinetic perspective reconciles life's apparent defiance of equilibrium: in effect, the fastest replicators persist without needing any guiding force. By analogy (albeit imperfect), one might speculate that neural and social feedback loops in humans enforce lifetime affect balance, similar to how chemical kinetics enforce DKS. Of course, any such analogy is limited – DKS concerns molecular populations and energy flux, whereas LoF concerns individual experience. But DKS underscores that far-from-equilibrium systems can obey strict persistence constraints purely through their dynamics. In short, even complex biological systems can exhibit lawlike balance underpinned by kinetics rather than teleology.

25.2.3 Feedback and homeostasis

Perhaps the most illuminating analogies come from control systems and homeostasis. The human body and mind are replete with feedback loops that keep variables within certain bounds (temperature, blood sugar, mood, etc.). A thermostat is a simple example: it activates cooling or heating to keep a room near a set temperature. Emotional

life could have similar regulators. In fact, psychology recognizes that we adapt to good and bad fortune – mechanisms like hedonic adaptation and opponent processes act as feedback that pulls us back toward a baseline. Classic homeostatic models emphasize returning to a set-point, while allostasis involves changing the set-point under different conditions (the body anticipates needs and adjusts its target levels). LoF, however, posits something beyond ordinary homeostasis: not only do we gravitate toward an emotional baseline, but the *cumulative deviation* from that baseline is actively managed such that it zeros out by the end. This sounds like a kind of “integral control” mechanism in engineering terms. In control theory, a regulator with an integral term doesn’t just respond to the current error (deviation from set-point), but also accumulates error over time and responds more aggressively if there’s a persistent bias. If our emotional system had an integral control component, it could conceivably work to eliminate long-term cumulative error (an emotional surplus or deficit) as the time horizon (lifetime) is about to close.

In Part VII of the book, we theorized exactly that: as one’s horizon shrinks (i.e. as one senses the approach of life’s end or end of a significant chapter), the mind might ramp up corrective actions. This would manifest as a strong drive to seek closure, reconciliation, or intense experiences that offset any long-standing imbalance. We see hints of this in real life – consider the oft-cited phenomenon of life review in the elderly or terminally ill, where people spontaneously reflect on unresolved issues and often attempt to make amends or find peace. It’s as if, knowingly or not, people have an internal gauge that says “time is running out, balance your books now.” On the neural level, one could imagine the brain increasing certain kinds of plasticity or emotional sensitivity that drive a person to address unfinished emotional business. This is speculative, but it offers a concrete framework: a negative feedback loop that grows stronger with age or approaching milestones.

We also note that any such system must work through our choices and behaviors. It might not be a separate mystical force, but embedded in our very instincts. For example, if someone has been riding high on success for years (large positive ledger), perhaps they subconsciously begin to take risks or feel restless, making choices that humble them, thus bringing their mood down to earth. Conversely, a chronically unfortunate person might eventually feel an urge to reach out, to try something drastically different, or might receive empathy from others that finally lifts them – thus a correction occurs. In prior chapters we described these processes as individuals co-producing balance: we strive, often without realizing, to regulate our feelings. The Law of Fairness would be the grand summation of all those small regulatory actions, plus the body’s physiological limits (e.g. you cannot cry forever, eventually exhaustion sets in and calm follows).

It's worth noting that while homeostasis and allostasis explain short-term and adaptive regulation, they don't automatically guarantee a zero lifetime sum. LoF would effectively require that beyond daily homeostasis, there is an overarching lifetime homeostat. Skeptics might call that fanciful, but it could be a natural extension – an organism that cannot in the long run self-correct might be less likely to thrive. Evolution may not “care” about perfect hedonic balance, but it does care about resilience and functional stability. A creature completely overwhelmed by pain or endlessly intoxicated by pleasure would be at a survival disadvantage. Thus, the argument goes, evolution might instill checks that prevent permanent extremes. LoF could be seen as the extreme limit of those checks – pushing not just toward viability, but all the way to neutral completeness.

We drew an explicit analogy that perhaps makes this clear: imagine programming an artificial agent with a reward system. If we wanted that agent to neither spiral into despair nor float away in euphoria (both states that could impair judgment), we might program a rule that as total reward or punishment accumulates, the agent's future rewards are adjusted to compensate. Engineers might implement this by, say, reducing reward sensitivity after a lot of rewards (making further pleasure harder to come by) and increasing reward sensitivity after a lot of punishments (making it easier to feel joy or relief). The result would be an agent that almost inevitably winds up near net zero reward at the end of its run. In Chapter 18, we described simulating such agents versus normal ones to see how their life trajectories differ. The LoF-like agents tended to avoid prolonged extremes and showed variance compression near the end – their total scores clustered near zero – whereas normal agents had a wider scatter of final scores. This thought experiment with AI provides a blueprint for how nature *could* implement LoF: through layered feedback mechanisms that dynamically adjust our “hedonic sensitivity” or motivation in response to our cumulative history.

To find evidence for this in humans, we can look at phenomena like the “aging positivity effect” and horizon-driven behavior changes (which we will discuss in a moment). But before that, let's incorporate an insight from measurement and information theory regarding these feedback processes.

25.2.4 Modeling and prediction

A rigorous way to evaluate LoF is to build mathematical models and see what they predict. In Part VIII, we outlined a simple stochastic model of life's emotional fluctuations. Imagine modeling life's ups and downs as a random walk – you have good and bad days somewhat unpredictably. Without any special mechanism, the net sum of such a random walk after many steps will typically wander away from zero (in magnitude), with some lucky individuals ending far positive, some unlucky far negative,

and most in between. Now, if we introduce a corrective bias that grows over time – akin to the integral feedback mentioned – the model changes. We specifically tested the idea of an Ornstein–Uhlenbeck (OU) process with a twist: an OU process normally pulls a variable toward a fixed mean (here, zero) with a certain strength; we let that strength increase as the remaining time decreases. Early in life, you wander somewhat freely (though there's still mild adaptation pulling you toward baseline), but later in life, the restoring force grows stronger to strongly bias trajectories toward baseline at the terminal time. This “horizon-weighted OU process” is a mathematical caricature of LoF.

Simulations of this model yielded distinctive patterns: as life nears its end, the variance of possible net outcomes shrinks dramatically – the trajectories seem magnetically drawn to zero at the final time. With a normal OU (constant resilience), you'd still get a spread of final totals (some lives end a bit positive, some negative, though not as wildly as pure chance). With the LoF-enhanced OU, the spread narrows dramatically; it's as if an invisible funnel guides paths toward neutral in the last stretch. We identified this as a clear empirical signature to look for: if LoF is true, the distribution of end-of-life integrated ledger totals across many individuals should be tightly centered near zero, much tighter than you'd expect from any known adaptation processes. Conversely, if we observe many people's integrated ledger totals and they show a wide variance, LoF in the strict form is falsified.

We don't yet have the longitudinal data over entire lifespans to perform this test conclusively – that is a project for future researchers, one we encourage by sharing our methodologies openly. But even in existing data, we sometimes see hints: some longitudinal datasets report a convergence in self-reported well-being in late life, despite very different life events. Also, extreme cases stand out precisely because they are extreme – they might be the few data points that remain off the zero mark, and if they truly never balance by the end, they are critical falsifiers. Thus, our model guides us to pay special attention to outliers: for LoF to hold, no outlier can persist to the end.

This modeling exercise underscores something important: LoF may sound like a vague philosophical idea, but it yields quantitative, testable outcomes. We turned it into equations and simulations, which means we can use the scientific method to probe it. By bringing in this systems modeling approach, we ensure that LoF is not just evaluated by anecdotes or intuitions, but by whether its predicted patterns match observed reality. In the spirit of systems science, it also forces us to clarify assumptions (e.g. how exactly does the “force” towards balance scale with time or magnitude of imbalance?). Making it formal either strengthens the case (if data align) or exposes flaws (if data contradict the model).

25.2.5 Information-theoretic considerations

Stepping into an information theory mindset, we can ask: *Why might a system (like a brain) benefit from balancing experiences?* One possible answer is avoidance of sensor saturation and maximization of information content. If an organism lived in unrelenting bliss, eventually its perception of bliss would flatten out – there's no contrast left to perceive finer gradations of happiness. Likewise, unrelenting pain could numb the senses over time – when everything is extremely bad, the difference between bad and slightly less bad might not register. Our sensory and emotional systems are most informative when they operate within a range that isn't maxed out on either end.

Thus, there's an argument that a finite nervous system may tend to favor variable experiences over being stuck at an extreme, because variable experiences carry more information and learning. A life that oscillates through highs and lows exposes the organism to a wider range of stimuli and internal states, potentially making it more adaptable and knowledgeable. In a sense, a perfectly imbalanced life (all one thing) could become *informationally inert*. By contrast, a life with balance has to experience both happiness and sorrow – it's richer in data. Some speculative thoughts in Chapter 17 floated this idea: maybe the brain automatically dampens sustained extremes because it gets more signal from change than from sameness. This aligns with known neuroscience: many neurons respond most strongly to changes or deviations from expectation, not to steady-state inputs. If you sit in a hot bath for a long time, the intensity fades as you habituate – your nervous system essentially says “no new information here.” The same might happen with emotional stimuli – a protracted period of one-sided emotion could lead the brain to attenuate that emotion's impact (reducing gain, so to speak) and seek novelty or contrast.

Another information-based argument we touched on is the idea of dynamic range adaptation. The brain has a limited “scale” to encode value or pleasure/pain. If life events push signals beyond the current scale, the brain might rescale. For instance, if you experience something unprecedentedly joyful, what once was a “10/10” happiness may now be felt as a lesser high because your internal scale expanded to accommodate the new maximum. Similarly, intense prolonged trauma might recalibrate your pain scale – minor annoyances no longer even register because your baseline moved. This constant recalibration would ensure that you continue to use the full range of your emotional capacity to discriminate new inputs. Over decades, such rescaling could mean that early extremes are remembered or felt in retrospect as not *quite* as extreme relative to later experiences. It's like compressing a lifetime of highs and lows into a finite bandwidth: outliers get compressed more, bringing the cumulative total closer to an even balance.

This is a subtle mechanism – more theoretical than empirically verified – but it’s plausible given how many adaptive coding tricks we know the brain employs to handle vision, hearing, etc. Emotions could be similar, always auto-adjusting such that we don’t get permanently stuck at one extreme of the dial.

In Part IV, when constructing the HCl, we implicitly addressed some information-theoretic challenges. We wanted a metric that captures as much *relevant* emotional information as possible without being swamped by noise or bias. The very act of measuring one’s hedonic state over a lifetime is an information compression problem: compress thousands of days of multifaceted experiences into one cumulative number. We approached this by carefully combining modalities (self-report, physiology, behavior) and calibrating them to be comparable across people and cultures. This cross-calibration ensures that our measure is robust – an important point because if LoF is a genuine invariant, it should hold regardless of culture or personality, meaning our units of measurement must be aligned. We cited cross-cultural studies and invariance testing to ensure that what we count as “zero” or “neutral” is meaningfully similar across individuals. This rigorous measurement approach is essentially applying information theory (maximizing signal, minimizing noise) to a very human question. It’s one of the strengths of our project: we didn’t settle for vague terms; we built an apparatus to track the “information” of a life’s emotional trajectory.

25.2.6 Interpersonal systems and network effects

While LoF is defined per individual, real individuals live in societies. No discussion of systems would be complete without considering the networked nature of human lives. People are not isolated nodes; we constantly influence each other’s emotional states through empathy, support, conflict, and culture. One might ask: even if there’s no *global* conservation of happiness (and we certainly do not claim that humanity as a whole has zero sum happiness at any given time), could social interactions help enforce each person’s balance? The idea here is that humans have evolved social feedback loops that *incidentally* promote balance. For example, when someone is deeply unhappy (a large negative imbalance), it often evokes concern and help from those around them – friends, family, or even strangers might step in to comfort, assist, or rescue. That influx of care can elevate the sufferer’s experience, moving them back toward neutral. In the opposite case, if someone is extremely fortunate or joyful to the point of excess, it can sometimes breed envy or social sanction, which might bring them down a peg (think of how communities react to boastfulness or how “survivor’s guilt” can temper one’s joy when others are suffering). While it might not be pleasant to consider, there is a kind of social levelling that happens informally: societies tend to frown on both extreme miserableness

(by feeling obligated to help) and extreme gloating happiness (by applying a bit of pressure or skepticism).

We discussed how cultural norms and stories reflect this balancing urge. Many cultures have proverbs akin to “don’t laugh too loudly, or you’ll cry later,” warning against tempting fate with excessive joy. Others emphasize helping those in need, essentially ensuring the lows are lifted. Even rituals around death and misfortune – such as communal grieving and support – act as balancing mechanisms, preventing individuals from getting indefinitely trapped in sorrow by surrounding them with care. In a network sense, information about one person’s imbalance (their suffering or their runaway success) spreads to others, and those others react in ways that tend to mitigate the imbalance. It’s an emergent social phenomenon, not a conscious conspiracy. Each person is simply following empathy or social emotion, but collectively it means extremely unhappy individuals often receive some joy from outside, and extremely happy individuals either share their joy (charity, generosity) or face modest pushback.

This doesn’t guarantee LoF – many people fall through the social safety nets or hide their struggles, and some elites manage to shield themselves from others’ leveling influences – but it contributes to a general trend. If one were to model society, you might find that when individuals are coupled in networks, their emotional trajectories influence each other in a way that reduces the overall variance of outcomes. In an extreme imaginary scenario, if the whole world emotionally *averaged out* through empathy, everyone would converge to the same state. Reality is far messier, but the concept of “fairness is networked” (a phrase we used in Chapter 22) recognizes that we often balance *each other’s* ledgers, not by design but through natural social dynamics.

Importantly, none of these social considerations change the individual definition of LoF – each person must balance internally by their own end. But it suggests that one factor in achieving that balance might be the presence of others. A completely isolated person might have a harder time balancing extreme experiences (no external help or hindrance), whereas someone in a rich social web has more corrective inputs available (friends can brighten a dark time; community can humble an inflated ego). This implies that LoF, if true, may rely on humans being social animals. It does not magically act on a hermit without any contact (that hermit might be an interesting test case!). In Chapter 19, when discussing “channels,” we emphasized keeping social and emotional channels open. Social connectedness is a major channel: it allows the import and export of emotional energy in a way that facilitates internal balance. If someone closes themselves off from all support (or all critique), they effectively cut a feedback loop, and their ledger might remain skewed longer or permanently.

Bringing the physical and systems perspective to a close, we acknowledge a dual lesson: the analogies have guided us to specific, testable aspects of LoF (like distribution of outcomes, presence of feedback behaviors), but also highlighted where the idea is speculative. Consciousness doesn't follow physical laws as straightforwardly as planets or pendulums. If LoF is real, it likely emerges from a complex interplay of biology, psychology, and social interaction rather than a simple equation. We used the language of physics and systems to clarify our thinking, not to claim we've proven a new law by analogy alone. The true validation or refutation of LoF will come from data – do we see those balancing feedback loops and end-of-life convergences in reality? Part of this final chapter's goal is to show we've done our homework in imagining mechanisms and consequences.

25.2.7 Where we go next:

The next part will shift from mechanisms to evidence and lived experience: psychology is where the rubber meets the road, because it deals with actual human behavior and mental processes that either realize or contradict the Law of Fairness. Now let's enter the rich domain of psychology and human experience, to see how the ideas discussed so far manifest in real lives and minds.

25.3 The Psychological Perspective

Having framed LoF in theoretical and mechanistic terms, we turn now to psychology – the arena of human experience, behavior, and mental processes. If the Law of Fairness holds, it must do so through psychological phenomena. In this section, we examine a range of well-established psychological concepts to see how they align with or challenge LoF. We will look at how people emotionally react to events (both immediately and over time), how they cope with extreme situations, and how individual differences (like temperament or cognitive biases) affect one's hedonic "ledger." We'll revisit classic ideas like hedonic adaptation (our tendency to return toward a baseline after good or bad events) and opponent-process theory (where any strong emotion triggers an opposite reaction later). We'll also delve into the evidence for human resilience and meaning-making in the face of trauma, as well as the sobering instances where people do *not* recover (chronic depression, PTSD, or suicidal despair). In essence, we ask: *Do people's lives usually show a drift toward emotional neutrality?* If so, how? And if not, what does that tell us about the limits of LoF?

25.3.1 Adaptation and opponent processes

One of the cornerstones of our argument has been that humans naturally adapt to changes in fortune. Psychological research often finds that after major positive or negative life events, people's happiness levels tend to move back toward their prior baseline over time. For example, in some studies of lottery winners, initial euphoria is followed by a return toward more ordinary levels of happiness over time. After an injury or loss, despair can follow, yet over months or years many individuals climb at least partway back toward their prior mood levels. This is known as the hedonic treadmill or adaptation. It doesn't mean everything is exactly as before – some events do cause lasting shifts – but the magnitude of lasting change is often smaller than people initially expect.

In the 1970s, psychologists Solomon and Corbit described the opponent-process theory of emotion: any intense emotional reaction (positive or negative) is accompanied or followed by a contrasting reaction that modulates the experience. For example, a skydiver feels terror during the jump and then an intense relief and elation afterward; over repeated jumps, the terror decreases and the post-jump euphoria also diminishes, stabilizing the overall experience. Or consider how after a stressful event, one might feel a wave of calm, or after a huge accomplishment, a sense of emptiness can set in (the "post-celebration blues"). These are short-term balancing acts where the body and mind seem to self-correct extreme states. We can view these opponent processes as micro-examples of what LoF posits on the macro scale. LoF essentially says: take this tendency

and extend it over the entire lifespan. Over long horizons, highs are counterweighted by lows and vice versa, such that by the end the integrated spikes and dips come out even.

Evidence for opponent processes is robust in controlled settings (drugs, thrills, etc.), but do they guarantee long-term neutrality? Not necessarily in each case – they ensure *diminishing returns* on extremes, which is part of the picture. Think of it this way: if every extreme joy you experience is followed by a relative crash, then a life with many joys is also a life with many after-joy crashes, which contributes to balancing the peaks. Conversely, a life with many sorrows often induces periods of relief, calm, or even post-traumatic growth where the person experiences a positive psychological change (greater wisdom, perspective, appreciation of life) because of the suffering. These observations align with LoF qualitatively: they show a push toward equilibrium. However, it's critical to ask if the equilibrium is usually *complete*. Traditional psychology would say no – not everyone fully bounces back from everything. There are people who suffer long-lasting trauma without equal positive rebound, and people who ride high on good fortune with only minor downturns. The law we propose doesn't deny these cases; it challenges us to see if, by life's end, even those cases find some late balance or if they truly remain unbalanced exceptions.

25.3.2 Extreme cases and resilience

The challenge for LoF is precisely those extreme scenarios. Imagine a child who suffers grave abuse and dies young – an intuitively horrifyingly “unfair” life with no evident compensation. Or someone who has a golden life of privilege and happiness and dies peacefully – an “unfairly good” life. LoF’s claim is that even these lives, if fully measured, would not be as lopsided as they appear. Perhaps the child had hidden moments of comfort, love, or internal resilience that outsiders wouldn’t guess, tipping the scales more toward neutral than we think. Perhaps the privileged individual had unseen struggles, emptiness, or existential dread balancing more of their joy than their smiling exterior reveals. These propositions sound almost desperate – and indeed they underscore how *audacious* LoF is. It asks us to reconsider whether our surface evaluations of a life’s happiness are incomplete. That is why in Part IV we advocated for multi-modal, continuous measurement (diaries, physiological monitoring, etc.) instead of relying on memory or external observation. If LoF is true, the data should show subtle comebacks and downturns even in lives we’d label as wholly tragic or wholly charmed.

Resilience is a well-documented phenomenon: a majority of people, even after severe adversity (bereavement, natural disasters, violence), eventually regain a decent level of functioning and even happiness. However, not everyone does. Some suffer chronic depression or PTSD that can last a lifetime. Under the Law of Fairness, these difficult

trajectories are not treated as exceptions that "escape" balance; they are cases in which the constraint (not a purpose) must still be satisfied by some lawful route by the death of mind—via ordinary channels, opponent processes, altered appraisals, dream-mediated counterweights, social relief, or (in the limit) horizon compression that neutralizes the ledger as time runs short. A central correction, developed across the book, is that we routinely undercount both sides of the ledger: the long, quiet seams of pleasure (a warm meal, lying down after hard work, bodily relief in the bathroom, ease with close friends, daydreaming, new love, sex, even the narcotic calm sought in substances) and the slow "background radiation" of pain (anxiety, shame, fear, illness, disappointment). A single day of deep ease can, in principle, weigh against weeks of toil, and a month of diffuse dread can, in principle, outweigh years of small delights when intensity and duration are weighted; LoF asks us to measure the stream as it is felt, not as it is narrated. These micro-sources and slow-burn costs/reliefs are precisely the kinds of mundane, non-miraculous ingredients LoF expects and tests for; balance, if real, is achieved through ordinary means and human interventions, not cosmic theatrics. Relatedly, the pattern many observe — front-loaded joy in youth and heavier burdens late in life — is not gratuitous cruelty on LoF but a familiar signature of endgame intensification: as horizons shorten, the system's shadow price rises, steering streams toward closure and making late-life compensations more likely to be strong, sustained, and salient (for good or ill).

The law is stated to hold for 100% of conscious streams, and it must apply even in extraordinary conditions. "Channels" name the plurality of lawful routes by which neutrality can be reached; they are not preconditions for the law's truth. If familiar channels close, the Queue System (QS) still operates by pruning what can line up next and by tilting felt options so that admissible policies remain those that permit compensability before the death of mind.

Now consider suicidal logic in the context of LoF. We begin with safety and ethics: LoF never justifies self-harm, and we should maintain strict standards for discussing suicide without glamorization, blame, or danger. This section is descriptive, not prescriptive, and the appropriate human response remains prevention, protection, and care. When someone takes their own life, it's often because they perceive their suffering to be unending or overwhelming, and any potential future joy hopelessly inaccessible. How would LoF account for this? Within formalism, a self-initiated terminal event is evaluated under the same closure criteria as any other terminal event; a verified end-of-stream imbalance counts against LoF rather than being explained away. In plain terms: LoF makes no moral claim about suicide, and it predicts only that the measurable ledger at terminal closure should still fall within the preregistered neutral bounds. Because this topic invites misreading, we reiterate: LoF is not a reason to die, and nothing in this theory

diminishes our obligation to keep channels open, expand care, and reduce pain. We should always include help-seeking pathways when teaching or publishing on these themes.

There is a poignant observation often made: some individuals who survive a suicide attempt (e.g., jumping from a bridge but surviving) report an immediate clarity and even regret as soon as they initiated the attempt, realizing that problems had solutions and life was worth living. LoF does not generalize from such anecdotes, but it can model them: a sudden opponent process or QS tilt can surface alternative policies once the immediate act is interrupted. That said, we resist romantic narratives and keep the focus on care, safety, and measurable streams. The logic here is sobering: LoF's promise of eventual balance is not a guarantee someone feels in the moment, and we should never expect a person in deep pain to "believe things will naturally even out." Psychological compassion dictates that we intervene, support, and actively create the conditions for balance (not assume some mystical force will save them). In research terms, any claim that a given suicide occurred "at balance" must be treated as empirically undecidable unless supported by composite evidence (e.g., HCI trajectories and documented channel access); absent such data, the case is not probative for or against LoF.

In sum, extreme cases test the limits of human psychological resilience. LoF does not rest on resilience being universal; it rests on a constraint enforced by mundane mechanisms (QS pruning and reweighting, opponent processes, cultural scaffolds), plus the fact that we underestimate both micro-pleasures and diffuse pains when narrating lives. Traditional psychology, grounded in observation, would counter that resilience is common but not universal—some people appear to succumb to adversity without equal rebound. That empirical tension becomes a measurement question in this book: we operationalize the stream with a composite index (HCI), predefine neutral equivalence bands, and specify QS falsifiers; if we find lives ending with sustained, uncompensated extremes under open channels and normal horizons, LoF fails. Two additional predictions sharpen the stakes: (1) because pleasures and pains accrue outside of salient "events," average lives should show moderate oscillations around neutral, with extremes rarer than folk memory suggests; and (2) as horizons shorten, compensations should intensify (including late-life illnesses or losses that, while never "justified," function as heavy weights in the ledger). Finally, on meaning-making: some readers will hear in QS's global pruning a blueprint-like feel that religious traditions describe as providence, karma, or the Holy Spirit's guidance. We do not adopt a teleology; we note the anthropological echo—cultures everywhere invent ledger metaphors and structured counterweights (penance, confession, charity) that mirror exactly what a

compensability-preserving controller would favor. The book treats these as cultural convergences, not proofs of the law.

25.3.3 Coping and co-production of balance

A key element often lost in philosophical debates is that human beings are active agents. We don't just passively receive happiness or suffering; we continually make choices to manage our emotional states. When hurt, we seek healing (if we can); when too giddy or unfocused, we might intentionally ground ourselves. There is an entire psychology subfield on emotion regulation and coping. Strategies range from reframing thoughts (cognitive reappraisal), seeking social support, exercising, distracting oneself, to less healthy ones like substance use or avoidance. Why do we regulate emotions? Fundamentally, to feel better or to stay functional. One could argue that all these micro-actions aggregate into the life-long balancing of the ledger. In earlier chapters we used the phrase "co-producing fairness" – meaning that if life does end up fair, it's not solely because of passive processes; it's because we and those around us *made* it fair through countless adjustments and acts of care.

Think of a person grieving a terrible loss. Under LoF, that deep pain must be balanced in the felt ledger before the death of mind. Balance does not arrive by magic; it is achieved through the ordinary channels people and cultures already use—therapy and, when indicated, medication; the steady presence of friends; sleep and dreams; spirituality and art; purposeful work and service; and the slow adaptations of time—while the Queue System (QS) prunes and tilts admissible options toward relief as horizons compress. The counterweight can take the form of restored peace, meaning, or joy after the loss; and in many lives the grief itself functions as the compensating weight for a long backstory of net ease and pleasure that had accumulated beforehand. Because we chronically undercount quiet, durable pleasures and likewise undercount diffuse pains, the offset is rarely obvious in narrative memory. A single day of deep ease can, in principle, balance weeks of toil; months of low grade dread can, in principle, outweigh years of scattered delights. LoF's claim, therefore, is not that some cosmic scale injects happiness, but that—under a universal constraint—the lawful mix of human agency, cultural scaffolding, and QS dynamics co-produce neutrality by terminal closure.

On the flip side, if someone experiences wild success and pleasure, do they automatically come down to earth? Often, reality intervenes – big success might bring new stresses or make previous joys less exciting. But also, people often *self-regulate* runaway pleasure for practical reasons: they know they can't stay on vacation forever, or they fear hubris, or they simply acclimate and start seeking what's next. In literature and wisdom traditions, there are warnings that too much pleasure can lead to downfall

(addiction, complacency, etc.), which essentially encourage proactive moderation. Again, human agency is key: we temper our highs as much as we soothe our lows (consciously or unconsciously).

I make this point to emphasize that LoF doesn't require believing in an unseen hand doling out pleasure or pain. It can be understood as the result of innumerable human decisions. If someone's "ledger" is deeply negative, they might consciously or unconsciously be driven to do things that add positives (reach out for help, change their environment, fight for improvement). If their ledger is sky-high positive, they might take steps (or life might present challenges) that bring it down some (responsibilities, empathy for others' suffering, etc.). In aggregate, these choices create a balancing trajectory. In fact, if LoF is true, it might simply be a grand way of saying: *humans strive for balance and generally succeed by the end*. That's a less mysterious formulation.

Where this becomes ethically salient (and we stressed this in Chapter 24) is that recognizing the role of choice means recognizing responsibility. If you see someone suffering, you cannot assume "oh, they'll balance out eventually." Rather, you should think "how can I open a channel for them to find relief?" because it might be through your action that their balance is achieved. Likewise, for yourself: believing in LoF is not about sitting back; it's about actively contributing to your own balance (seeking meaning in hardship, staying humble in fortune, etc.). In psychological terms, this aligns with many positive practices: gratitude exercises (to not take happiness for granted), grief work (to process and release pain), forgiveness (to let go of bitterness), and so forth. All these practices smooth out the peaks and valleys, nudging life toward equilibrium. The difference with LoF is the assertion that these nudges, over a lifetime, will always suffice to complete the job.

25.3.4 Growth from suffering and meaning-making

One of the more encouraging findings in psychology is that suffering often isn't in vain – people frequently report positive transformations following adversity. This can include a deeper appreciation for life, improved relationships, newfound personal strength, spiritual development, or clarity on what matters (phenomena often grouped under post-traumatic growth). Viktor Frankl, a Holocaust survivor and psychologist, famously wrote about finding meaning through suffering; many therapists build on this, helping patients reframe trauma as something that, while not wished for, can catalyze growth or purpose. In the context of LoF, such meaning-making is a clear example of a compensatory positive emerging from a negative. The scales might not be balanced in a hedonistic sense (pain is pain), but if the person feels that their suffering led to something good –

wisdom, empathy, a life calling – that adds to the positive side of the ledger in a profound way.

We have been careful in the book not to slip into glib statements like “suffering is good for you” or “everything happens for a reason.” We categorically reject the idea that severe pain is justified because some happiness will follow. Instead, our focus has been on acknowledging the pain and then actively seeking or allowing whatever good can come in its wake. LoF, if couched in existential terms, suggests that life might inherently supply each person with the ingredients for a meaningful narrative that, at the end, they can accept as complete. That often involves forgiveness, creativity, love, and reconciliation – all processes by which people take the bad and integrate it into a larger story that has positive value.

Take, for example, someone who endured injustice but later becomes a champion for others facing similar struggles. Their past pain, while not erased, is balanced by the purpose and positive impact they derive from it. Or consider an elderly person who has had multiple heartbreaks and losses, yet in their twilight years, they often express peace, saying things like “I’ve made my peace with what happened, and I cherish the good that also came.” Hospice workers and gerontologists note that many people (not all, but a notable many) reach a state of equanimity near life’s end, even if their life was hard. They reconcile with estranged family, they forgive old enemies, they find comfort in faith or philosophy, they impart wisdom – all hallmarks of tying up the emotional loose ends. It’s as if, given the chance, people have an innate drive to complete their emotional journey in a satisfactory way.

In LoF terms, final balancing often occurs via appraisal and meaning-making that feed back into the hedonic ledger, not only through raw sensation. Pain is not repaid unit-for-unit; it can be counterbalanced by meaning, growth, contribution, or coherence, while pleasure is tempered by perspective and responsibility. This does not redescribe suffering as good or assume proportional payback. It describes how, in global life review, people often judge that the difficult parts and the good parts carry comparable weight in who they became.

Methodologically, HCI targets first-order affect (the moment-to-moment stream). Narrative appraisal is a second-order judgment about the life as a whole. By the end of life, many people rely more on whether their story hangs together and feels worth it than on a count of happy versus sad days. Under LoF, fairness at the narrative level means the person feels broadly at peace—neither triumphs nor losses dominate their identity—because both are integrated into a coherent life story. This is a tendency, not a guarantee; severe constraints (e.g., illness, trauma, structural harms) can block integration.

25.3.5 Subjective bias and fairness

One complicating factor in all of this is that *different people naturally interpret their experiences differently*. Two individuals could live through very similar circumstances and yet end with very different hedonic ledgers because of their mindset. A person with a sunny disposition might count every little blessing and shrug off negatives, ending up feeling their life was mostly happy; a person with a pessimistic bent might do the opposite, focusing on slights and disappointments even amid fortune. Does LoF account for such differences? It would have to. The law cannot rely on an external observer's tally; it must be true in each person's own frame of reference. This means that subjective perception is key: the optimistic person's ledger and the pessimistic person's ledger are calculated according to their own internal scales and interpretations.

LoF also allows balancing via shifts in appraisal. Someone who habitually discounts positives can move toward equilibrium not because the world changes, but because their weighting does—through perspective-taking, gratitude, or late-life reappraisal that recognizes goods they previously ignored. Conversely, someone who overweights positives may rebalance by acknowledging costs and harms they had downplayed. The mechanism here is not cosmic correction but cognitive and social recalibration over time. In HCI terms, the event stream may be unchanged while the aggregation weights shift. Again, this pathway is common but not universal.

Psychological science provides tools to measure these biases (like questionnaires for dispositional optimism, depressive realism, etc.). Incorporating those into LoF tests is important. We mentioned in our methodology that calibration is individualized. That means if someone is a curmudgeon who rates even a decent day as “5/10” happiness, our HCI would capture their *relative* ups and downs around their personal baseline, not judge them by someone else's scale. LoF would then say that even that curmudgeon will have an equal mix of relatively good and relatively bad experiences by their own judgment. It might be that their “good” never goes above what another person calls mediocre, but internally, their peaks and valleys should balance. This is a tricky point: fairness doesn't mean equal absolute happiness across people; it means each person tends to use the full range of *their own* emotional capacity so that weighted highs and lows balance over time. Someone dispositionally low might live oscillating between flatness and despair; LoF does not predict large euphoric highs for them, but it does predict recurrent relief, numbness, small comforts, or brief satisfactions that, within *their* range and when properly weighted for intensity and duration, bring the ledger toward balance enough times.

The existence of subjective bias also means verifying LoF is challenging – we need to measure *within* each person accurately. But if it holds, it would be quite a remarkable assertion of psychological universality beneath diversity: no matter your personality, your cognitive biases, or your lot in life, you will experience enough of your subjective positives and negatives to call it a draw. We often think some people are just happier overall than others (and that's empirically true in studies – happiness has trait-like stability). LoF would controversially claim those differences are superficial or transient, and that by life's end, even the grouchiest and the cheeriest might have more equal totaled experiences than anyone would have guessed. This is one reason psychologists might resist LoF: it seems to undercut the idea of stable individual differences in well-being. It pushes a kind of hidden equality.

25.3.6 The final stage – aging and acceptance

If LoF exerts any influence, we would expect to see its effects most clearly as people approach the end of life. And indeed, there are intriguing findings about aging that align with a drive toward emotional balance. Research shows that older adults, on average, experience less emotional volatility and often report equal or greater well-being than younger adults, despite the physical and social losses that come with age. This is sometimes called the paradox of aging: objectively, things might be worse (health issues, friends passing away), yet subjectively many older people are more content. One theory explaining this is socioemotional selectivity theory (by Carstensen and colleagues), which proposes that as people perceive their remaining time getting shorter, they prioritize what matters most – typically emotionally meaningful experiences and relationships – and let go of peripheral or negative pursuits. In practice, older individuals focus on close loved ones, hobbies that bring joy, and often avoid needless conflict or stress. They also tend to process information with a positive bias: studies find that the elderly pay more attention to positive stimuli and are more likely to remember positive over negative events, a pattern dubbed the “positivity effect” (Mather and Carstensen, 2005).

Neurologically, some imaging studies report decreased reactivity of the amygdala to negative images and relatively maintained responses to positive images in older adults, compared to younger adults. In addition, older adults can show greater engagement of frontal brain regions during affective tasks, which is consistent with increased reliance on regulation; however, these measures are correlational and do not, by themselves, establish a causal mechanism. All this points to a shift toward emotional homeostasis in later life. It's as if the brain gradually tilts the playing field to favor contentment – possibly reflecting motivational changes as the end draws near.

Culturally, we see convergence as well. In many cultures, there's a strong ethos of seeking closure in one's final years: making amends, reflecting on life, experiencing generativity (like sharing wisdom with younger generations). Hospice care practices around the world – whether informed by religion, tradition, or modern psychology – invariably emphasize comfort, reconciliation, and unburdening the dying person of regrets or pain. Loved ones gather to express gratitude, forgiveness, and love. Old conflicts are forgiven. People often say things like “nothing else matters now except that we are here together.” These behaviors quite literally help balance the emotional books: they reduce fear, guilt, and regret (negative feelings) and amplify feelings of love, gratitude, and spiritual peace (positive feelings). We might view this as society's intuitive realization that a person should ideally meet death in a neutral or positive state of mind.

It is compelling that both individual psychology and collective practices spontaneously orient toward balance at life's end. While not proof of LoF, it's consistent with it. It's as though human beings have an innate or social “end-of-life balancing protocol.” When channels are open – meaning the person is cared for, pain is managed, and they have the opportunity for emotional resolution – many people die with a sense of contentment or at least without overwhelming anguish. This doesn't happen for everyone; some tragically die in pain or mental turmoil. But noticing how *often* people find peace in the final stage, even those who struggled much of their lives, is suggestive that there are convergent forces at work.

For example, a person who had a very joyful youth and midlife but then faces illness and dependency in old age might feel their dignity and happiness ebb – a late-life suffering that balances earlier bliss. Conversely, someone who had a rough start and midlife might find their later years unexpectedly calm – perhaps retired from stress, enjoying grandchildren, reflecting proudly on having “made it through,” thus experiencing a late-life contentment that compensates for earlier woes. These are anecdotal patterns, but they resonate with many of us. We often observe that “life has a way of evening things out.” The cheerful war veteran who endured unspeakable battles gets a tranquil old age surrounded by family; the once-spoiled celebrity who had it all experiences loneliness and health struggles at the end. While not a rule, such stories are common enough to keep the LoF question alive.

25.3.7 Summary of psychology – fairness on trial

From a psychological perspective, the Law of Fairness is both enticing and exasperating. It is enticing because it ties together many threads: adaptation, resilience, coping, growth, and aging, painting a picture that all these human capacities collectively ensure an equilibrium. It's a grand tapestry that, if true, would highlight an extraordinary

robustness in the human spirit: that no matter what, we have the mechanisms to end up *okay*. It's exasperating (especially to scientists) because it goes beyond the evidence in insisting this holds with no exceptions. Psychology teaches us that while most people show resilience, some do not; while most eventually move on, some remain stuck. It cautions us about survivorship bias – perhaps we hear more from those who find meaning in suffering than from those who are quietly defeated by it. The “no exceptions” clause of LoF sets a high bar: a single clear counterexample (with thorough measurement and support available) calls the law into question.

Throughout this book, I've tried to address these concerns. For instance, we have specified that, for LoF to be taken seriously, we must look at cases with intact channels—the ordinary pathways through which both pleasure and relief can occur (e.g., food that tastes good, rest after hard work, bodily comfort, time with close friends, sex or affection, daydreaming, small wins), alongside the pathways of pain (e.g., anxiety, illness, frustration). People routinely miscalibrate both sides—often underestimating the intensity and duration of everyday positives and also the drag of chronic negatives—so any test of LoF has to correct for these biases. If someone truly never had access to compensatory channels—say they lived and died under relentless harm with no realistic respite—then, within this book's stated scope, that would count as evidence against LoF rather than a reason to revise the scope.

To avoid “moving the goalposts,” falsification must be framed at the lifetime level, not a late-life snapshot. Thus, if—despite intact channels and appropriate bias-corrections—someone's integrated, longitudinal assessment (with intensity weighting) remains overwhelmingly negative up to the end *and* there is no credible evidence of earlier surpluses sufficient to offset late-life suffering, that would count against LoF. Conversely, if rigorous longitudinal studies show that—even among those facing immense hardship—the integrated ledger tends toward balance (with intense joys at times offsetting long stretches of effort or pain, as in childhood/young-adult surpluses later counterbalanced by illness or loss), that supports LoF's plausibility.

A vital point we reiterated is that LoF is not a prescription to be passive or naive. It would be dangerously easy for someone to misunderstand and think “everything will just work out, so I need not act or worry.” That is *not* what the evidence of psychology suggests. Things work out (when they do) *because* we take action: seeking therapy, leaning on friends, learning from pain, restraining our excesses, and so forth. And sometimes things don't naturally work out without intervention—which is why we must intervene. If you truly believed LoF in a fatalistic way, you might not rush to help the distraught friend (assuming they'll bounce back eventually), or you might avoid joyous experiences (fearing future

pain as payback). Those would be misinterpretations. We have repeatedly emphasized a proactive stance: use LoF as motivation to foster balance, not as an excuse to ignore imbalance. Knowing (or hoping) that no one is beyond eventual repair means we should double down on providing the tools and support for that repair – not leave it to chance. In psychological practice, this aligns with encouraging people to have hope *and* to take steps toward healing.

In conclusion of the psychological perspective, we can say the following: Human minds exhibit many balancing forces – from automatic emotional processes to deliberate coping efforts – which lend credence to the idea that lives gravitate toward equilibrium. Most people's emotional trajectories do show a tendency to level out extreme highs and lows over time. Whether this tendency is an ironclad law or just a strong trend with exceptions remains to be seen. If anything, posing LoF has been a fruitful provocation to psychology: it compels us to look at life narratives in full, to gather data on complete lifespans, and to identify the factors that promote or impede emotional recovery. Already, the journey of exploring LoF has illuminated how vital things like social support, adaptability, and meaning-making are in shaping lifetime well-being. If LoF is true, it's a tribute to human resilience and the subtle design of our emotional systems. If it's false, delineating its failure will still teach us exactly where and why some lives remain unfair – and that knowledge can guide targeted interventions (for instance, identifying that maybe certain mental illnesses or social conditions break the balancing mechanism, which tells us where to focus resources). In that sense, psychology “wins” either way: we either discover a unifying law or we map the limits of human resilience with greater precision than before.

25.3.8 Where we go next:

Having weighed the evidence and intuition from psychology, which largely support a balancing *tendency* but debate its universality, we now step into a different arena. In 25.4 we explore the realm of spiritual and moral philosophy. Humans have long grappled with fairness not only through science, but through religion, ethics, and storytelling. The final perspectives we'll explore connect the Law of Fairness to those rich traditions, asking how this hypothesis mirrors or diverges from what our spiritual heritage and moral consciousness tell us about why we suffer and rejoice.

25.4 Spiritual and Moral Parallels

The idea that life might be ultimately fair is not new – it has deep roots in religious and spiritual thought. What is new is applying fairness to total felt experience and stripping that idea of divine intervention or moral judgement. Nonetheless, in this section we venture into comparative philosophy and theology to see how LoF resonates with or challenges traditional beliefs. We will examine parallels with Eastern concepts like karma and yin-yang balance, and Western religious themes such as divine providence or the notion that “God gives each person only what they can bear.” We’ll consider metaphors like life as a blueprint or script designed with an allotted balance of joy and pain – metaphors that invite speculation about destiny and free will. We will also discuss the interplay of LoF with moral values: how the daily “ledger” of good and bad deeds (sins and virtues) might relate to the hedonic ledger of experiences. Does living ethically influence the balance of happiness and suffering, or is that a separate account altogether? And how do religious frameworks like the concept of the Holy Spirit (viewed as a comforter and guide) or the principle of cosmic justice interpret something like LoF? By exploring these questions, we acknowledge that while our thesis is scientific, it touches on age-old human yearnings for fairness, meaning, and justice – concerns traditionally addressed by theology and moral philosophy.

25.4.1 Echoes of karma and divine justice

One cannot discuss an ultimate balancing of life’s experiences without invoking karma, the ancient Indian concept that one’s actions (and even intentions) cause future happiness or suffering, ensuring a just balance over time. At first glance, LoF might sound like karma’s twin: “everyone gets what’s coming to them in terms of pleasure and pain.” But there are crucial differences. Karma is fundamentally a moral doctrine – it ties your good or bad experiences to your good or bad actions (often carried across lifetimes). LoF, in contrast, is morally blind: it doesn’t propose that suffering is punishment or happiness is reward; it simply says both will happen in equal measure. In a sense, LoF is a kind of karma without morality or without transmigration – all contained in one life and not concerned with whether you deserved the outcomes. This makes LoF less about justice in the ethical sense and more about a neutral natural *lawfulness*.

Yet, for someone steeped in karmic belief, LoF might appear as a secular reformulation of the same cosmic principle. Instead of cosmic ledger-keepers tallying merits, we have psychological and biological processes maintaining equilibrium. Some religious thinkers might even welcome LoF as scientific validation of a karmic universe, minus the reincarnation element. It’s important, however, not to conflate the two: under karma, a vicious person could have a pleasant life if their past-life merits were good, but eventually

(maybe in another life) justice catches up. Under LoF, that same vicious person must experience a balance *within this life*, but not because of their vice – just because that's how experience unfolds. In fact, LoF can be unsettling from a traditional religious view because it implies even a cruel person will have as much joy as sorrow by the end (which offends our sense of moral justice), and even a saintly person will suffer as much as they're happy (which seems to undercut reward for virtue). In this way, LoF is closer to the notion of "rain falls on the just and unjust alike" – it's impartial and doesn't distinguish saint from sinner in dispensing life's fortunes.

Western religious traditions often wrestle with the idea of *theodicy*: why a good God allows good people to suffer and bad people to prosper. Various answers are given: it's temporary (justice in heaven or later), it's a test or soul-building, etc. LoF offers a bold proposition: maybe by the end of life, even if it looked uneven in the middle, everyone's experience is balanced. This would be a sort of built-in theodicy – not necessarily a satisfying one to those who expect virtue to be rewarded, but a comforting one in that *no one is utterly forsaken or purely fortunate*. For a religious mindset, one could interpret LoF as evidence of divine mercy or design: perhaps God ensures that each person, regardless of their moral status, gets equal shares of life's sweetness and bitterness – not as justice for their deeds, but as part of the soul's journey or a compassionate cosmic plan. For instance, Christian theology talks about God's grace and the idea that burdens are tailored to what one can bear. LoF resonates with the saying often attributed (perhaps apocryphally) to scripture: "God will not give you more than you can handle." In our framework, that translates to: everyone's total burden (suffering) will be matched by their strength and solace (happiness) so that it's never imbalanced beyond what they can carry to the finish.

An interesting theological parallel is the concept of Providence – the belief that there's a divine plan guiding events toward good ends. LoF is like a secular Providence: it doesn't say every event is good, but that the sum total of events yields a neutral completeness. Some might see that as a faint echo of divine Providence making sure each life is, in the end, "complete" or "fair" in experiential terms. However, where Providence or karma usually imply intention (God's will or cosmic justice), LoF implies a mechanism (nature's equilibrium). This highlights a theme: LoF can be a Rorschach test for spiritual interpretation. A devout person might say, "Yes, this is how God secretly works – not rewarding or punishing based on our small view of justice, but giving each soul a full measure of life's depth." A skeptic might say, "This has nothing to do with God, it's just biology and luck evening out." The convergence is that both see a pattern of balance; the divergence is in the source of that pattern.

25.4.2 The blueprint of life – destiny and choice

Many religious and spiritual traditions hold that each person has a destiny or a life plan. In some esoteric beliefs (and even in some interpretations of quantum metaphysics), there's the idea of a life script or blueprint that outlines the major events one will face. Often, these plans are thought to include certain challenges and blessings – almost like a plot that ensures the character (the person) goes through a meaningful arc. The blueprint metaphor for LoF would be: imagine that before birth (or by nature's design), each soul is assigned a blueprint that contains equal amounts of joy and sorrow, arranged perhaps in different sequences. One person's blueprint might give them a happy childhood and tough later years; another gets hardship early and peace later; a third alternates trials and triumphs throughout. There may be branching pathways – points where choices can send you down different routes – but all routes converge to the same net balance at the end. This is a fanciful way to visualize LoF, but it resonates with the observation that some lives seem front-loaded with pain or pleasure and then switch toward the opposite in later chapters.

If life were such a blueprint, it would reconcile free will with predestination in an interesting way: you have freedom to choose your path, but all paths have to obey the “area under the curve” rule of zero net. It's like being allowed to wander on a landscape that rises and falls (happiness and sorrow hills and valleys), but no matter which route you take, the total ups and downs encountered will sum to the same by the end. This view could comfort those who believe “everything happens for a reason” – except here the “reason” is not a moral one but a structural one (to maintain balance). It also puts a new spin on misfortune and luck: maybe those who suffer more at one stage are simply using up more of their “sorrow allotment” then, and thus might expect more reprieve later, and vice versa. Indeed, everyday talk sometimes reflects this intuition: “She's had all her bad luck early in life; I hope the rest will be smooth,” or “Things have been going almost too well for him; life has a way of throwing a curveball eventually.”

From a spiritual perspective, one could imagine a creator or universal intelligence setting up the world this way out of a sense of equity – not moral reward, but experiential richness. Perhaps the goal is that every soul experiences a full range of what life has to offer, so none only taste sweetness or only bitterness. This ties to the idea that perhaps the purpose of life is experiential rather than punitive or remunerative. In some mystical traditions, God (or the Universe) experiences itself through each of us, and LoF would ensure that each life contributes a balanced diet of experience to the whole. This is, of course, speculative and well beyond what our empirical approach can support, but it's a poetic interpretation that some readers might find appealing.

The blueprint idea also highlights the role of choice. LoF doesn't mean everything is fixed regardless of what we do. It could be that our choices determine *how* we encounter our quota of joys and sorrows. Take two individuals with the same "balance quota": one might achieve it through mild oscillations (making cautious choices, they avoid extreme highs and lows but still accumulate smaller pleasures and pains that add up), while another might live wildly (having ecstatic highs and crushing lows). Both end neutral, but one took a rollercoaster and the other a country road of gentle hills. This suggests a moral dimension: even if the net is fixed, we might prefer certain ways of getting there. Perhaps a virtuous life doesn't spare you suffering, but maybe it spares you certain kinds of suffering (like the suffering of guilt or the consequences of hurting others), replacing them with different sorrows (maybe the sorrows that come with sacrifice or empathy, which some might argue are "nobler" sorrows). Likewise, a vicious life might not buy more happiness net, but it might front-load superficial pleasures followed by deeper agony (like regret, isolation, infamy). Thus, even if LoF holds, *how you live still profoundly influences the texture and timing of your joys and sorrows*. This is where secular fairness meets moral framework: being good might not give you *more* happiness, but it might give you a kind of happiness you can feel at peace with, and a kind of suffering you can endure with dignity. Being bad might give you thrills that are fleeting and sufferings that cut to the soul. In this way, the daily ledger of virtue connects to the hedonic ledger – not by altering the sum, but by altering the path of reaching that sum.

25.4.3 Virtue, sin, and the hedonic ledger

All major religions and ethical systems emphasize that our actions have consequences for our own soul or well-being, not just for others. Modern psychology often echoes this in terms of mental health: for instance, chronic guilt or shame (from doing things one knows are wrong) can create a burden of suffering, whereas acts of kindness can bring a lasting warmth or sense of peace. How might everyday sin and virtue factor into LoF? If someone lives kindly, generously, and with a clean conscience, they may still face external hardships (accidents, illness, loss – because LoF isn't moral). But internally, they probably carry less emotional turmoil; their suffering may come more from outside events than inner demons. A cruel or selfish person might enjoy external advantages (perhaps stepping on others to get ahead, indulging in pleasures without regard), but internally they might accumulate stress, paranoia, emptiness, or the weight of guilt. These internal states contribute to the hedonic ledger significantly: guilt and loneliness are sufferings; inner peace and self-respect are pleasures (or buffers against pain).

In a **LoF** framework, virtue and vice do not change the **requirement** of balance, but they change what fills the balance sheet. A virtuous life may contain fewer self-inflicted

negatives; a vicious life loads the negative side with self-created pain (guilt, fear of retribution, lack of genuine love) even if circumstances look good. Many spiritual teachings warn that the worst suffering is a tormented conscience. LoF would predict that a tyrant who “has it all” is not ahead in net happiness because internally they carry misery. History and literature reflect this pattern: the tyrant dies paranoid and unloved; the greedy rich feel empty. Conversely, saints may suffer socially or physically yet report inner joy or peace.

In Christian theology, there’s the concept of the Holy Spirit as a presence that gives comfort, guidance, and conviction of sin. One could metaphorically relate this to LoF’s mechanism: the Holy Spirit “balances” believers internally by consoling them in pain (adding to the positive side) and pricking their conscience when they stray (introducing necessary discomfort to correct course). This is obviously a theological interpretation, not something our research tested, but it shows how religious people might map LoF’s balancing to divine action in their lives. They might say, “When I was too proud and happy, God humbled me with challenges; when I was broken, God sent me comfort – the Lord giveth and taketh away, blessed be the name of the Lord,” expressing a faith-based LoF in effect.

The notion of sinful pleasures is also interesting here. Many religions caution that illicit or excessive pleasures come with a price – either in this life (hangovers, addictions, consequences) or the next. LoF would concur that every spike of pleasure *will* be paid for, but not as punishment, simply as balance. If someone binge-indulges in something, they often feel a crash or emptiness after. That’s an opponent process in psychological terms, but in moral language it might be framed as the wages of sin. The difference again is LoF doesn’t moralize it; even wholesome pleasures will be balanced by some pain (which might be wholly undeserved). However, if one believes in a just universe, one could overlay a moral interpretation: perhaps the pain that follows a pleasure is sometimes directly tied to whether that pleasure was righteous or not. For instance, pleasure from helping others might be balanced by later sorrow that has nothing to do with that act (just life being life), whereas pleasure from exploiting others might be balanced by later sorrow that is a direct consequence (like people turning against you). Both end up balanced, but the latter carries a sense of justice.

What about collective sin or virtue? Does LoF apply beyond the individual? Our formulation says no, each individual is their own closed system. But spiritually, communities often believe in shared karma or collective punishment/merit. While we haven’t endorsed any global fairness law (we explicitly avoided that because clearly some generations or groups suffer far more than others in history), one could speculate

that maybe across the grand sweep of existence, even societies or the world might seek an equilibrium. This ventures into theological territory of divine justice across nations or eras. LoF isn't equipped to handle that—it's beyond our evidence and likely not true in any straightforward sense (history is rife with unfairness that wasn't later compensated on a social level; often it was *other people* who eventually compensated or rectified injustices). So we won't assert a collective LoF. But an individual with a strong moral worldview might personally reconcile with LoF by thinking of it like, "I will suffer for my sins and be comforted for my good deeds, somehow, by life's end." That's not exactly what we claim scientifically, but it might be a useful personal ethos if it encourages good behavior and acceptance of hardship.

25.4.4 Collective balance and exemplary individuals

Finally, it's worth touching on a quasi-spiritual notion: some individuals seem to carry extraordinary loads of suffering or blessings, almost symbolically. Think of figures like saints, martyrs, or, conversely, extraordinarily fortunate people. In religion, sometimes one person's suffering is seen as redemptive for others (the ultimate example being Christ's suffering in Christian theology). While LoF is strictly individual, one might wonder: do some people's lives tilt one way because they are part of a larger balancing act across humanity? This is not something we can prove or disprove, but it's a question that arises when seeing gross disparities. Perhaps one deeply suffering person triggers compassion and goodness in many others, thereby distributing the balance indirectly (their suffering yields positive emotional growth in others, sort of spreading the balance around). Or one extremely joyful, inspiring person (like a blissful guru or a philanthropist) might absorb others' sorrows or inspire actions that alleviate suffering in the community, thus sharing their "excess" happiness.

This kind of poetic viewpoint is reminiscent of the concept of bodhisattvas in Buddhism – enlightened beings who delay their own final liberation to help balance the suffering of others. It also resonates with the Christian idea of bearing one another's burdens. In a universe with LoF, perhaps those who have more capacity or strength end up, by choice or fate, carrying more suffering to lighten the load on others, and those who have been given much joy spread it around. Our hypothesis doesn't formally accommodate this, but spiritually-minded readers might see in LoF a call towards solidarity: if life will balance out, maybe we are meant to be agents of that balancing for each other. The earlier section on societal implications indeed urged that – we should actively smooth out each other's extremes.

When we examine exemplary individuals – say, a Mother Teresa who suffered with the poor but also radiated a profound joy, or an Adolf Hitler who inflicted suffering massively

and ended in misery – they become almost allegorical evidence that extremes do circle back (one in a virtuous cycle, one in a vicious cycle). The saint takes on suffering but gains spiritual happiness; the tyrant amasses power and pleasure but ends in a bunker of despair. These stories are often used as moral lessons, but they also fit LoF’s narrative in a karmic way: not because of cosmic punishment, but because of how these lifestyles inevitably play out on the human psyche and in social response.

In drawing these parallels, we must emphasize: The Law of Fairness as presented in this book is not a moral law. It doesn’t claim the universe rewards good and punishes evil. People looking for that kind of justice must look to philosophy, religion, or human courts. LoF is agnostic to moral desert. Yet, by sheer happenstance or perhaps deep design, if LoF holds, it ensures a kind of existential equality: everyone’s life, saint or sinner, rich or poor, is equally a mix of sunshine and rain. There is something profoundly democratic about that – it would mean that at the level of felt life, no one is inherently better off than another in the grand total. This might offend those who feel the virtuous *ought* to be better off, but it also provides a humbling perspective that every human experience is, in sum, equivalently complex and filled with both joy and sorrow. It might engender empathy: the one you envy for their happiness has or will have their sorrows, and the one you pity for their sorrows has or will have their joys. We are all, as the saying goes, fighting great battles and enjoying secret triumphs behind the scenes.

Bridging spiritual and moral wisdom with our hypothesis enriches the narrative but doesn’t prove anything. It does, however, show that the intuition of balance is deeply embedded in human culture. Perhaps our scientific exploration of LoF is part of a timeless human endeavor to understand fate and justice, now couched in the terms of data and systems rather than scripture and myth. As we close this philosophical and spiritual reflection, we carry forward a sense of awe: if LoF is true, it reveals a hidden thread of connection between empirical reality and the perennial philosophy of balance. And even if it is not literally true, it aligns with practices and attitudes (moderation, compassion, hope) that spiritual traditions have long advocated as pathways to a good life.

25.4.5 Where we go next:

We head back down to earth, looking at how these ideas play out in our collective lives and what embracing (or refuting) LoF means for society. We will discuss practical ethics and communication – how to handle the concept of LoF in the public sphere without distortion, and how society might change if it took this hypothesis seriously.

25.5 Societal and Ethical Implications

The Law of Fairness, if engaged with seriously, has implications far beyond individual lives. It challenges how we as a society view luck, misfortune, success, and suffering. In this section, we explore how different cultures might react to the idea of LoF and what ethical considerations arise in communicating and acting on it. We consider the risk of fatalism or complacency – the dangerous idea that “since everything balances out, we need not strive to help or improve.” We then outline how, properly understood, LoF could inspire social policies and community behaviors that emphasize healing, second chances, and empathy. We also address the importance of responsible communication of a concept as potentially sensational as “life is fair.” If miscommunicated, it could be trivialized or weaponized; if communicated with nuance, it could encourage long-term thinking and mutual care. Lastly, we paint a picture of a society that embraces LoF’s insights: how healthcare, justice, education, and interpersonal relations might subtly shift to align with the goal that *no one’s story ends in unresolved imbalance*. This is speculative, but it helps illustrate why this whole exploration matters – not just as an academic question, but as something that touches on human dignity and how we treat one another.

25.5.1 Cultural perspectives on fairness

Reactions to LoF are likely to vary across cultural lines. In cultures influenced by Eastern philosophies (Hinduism, Buddhism, Taoism), the notion that life seeks balance may feel intuitive. The concept of yin and yang – opposite forces in harmony – is deeply embedded in Chinese thought, for example. A Taoist reading of LoF might be, “Yes, for every joy there is a sorrow, for every sorrow a joy; the wise person flows with this natural rhythm without excessive attachment or aversion.” Buddhism’s emphasis on the middle path and the impermanence of both suffering and pleasure also resonates: LoF can be seen as anicca (impermanence) writ large – neither happiness nor suffering is permanent, each will give rise to the other, so one should seek enlightenment beyond both. Thus, Eastern audiences might nod in recognition, though they might also caution that true liberation is not in the balancing of samsara (the cycle of worldly life) but in transcending it.

In Western, especially more individualistic cultures, LoF might rub against ingrained ideas of meritocracy and the Protestant work ethic (“you get what you earn”). Westerners often subscribe to the just-world hypothesis – a psychological bias where we assume if something bad happened to someone, they somehow brought it on themselves (because we want to believe the world is just). LoF directly contradicts a moral just-world: it says good people will suffer too and bad people will also have happiness. That can be uncomfortable. It also goes against the secular belief that through effort and control we

can maximize our happiness. Many Western self-help paradigms promise that you can conquer adversity and “live your best life” full of happiness by making the right choices. LoF says: no matter what you do, some sadness will find you; no matter how you optimize, you will not outpace the balance. That could be seen as pessimistic or limiting in such cultures. On the other hand, Western religious folk might see it as affirming a mysterious form of divine justice (as discussed earlier).

In Chapter 21, we discussed how to frame LoF depending on the audience. For a religious community, one might frame it as “No one suffers beyond their measure – there is a merciful balance in each life.” For a secular humanist crowd, we’d emphasize “We all share a common humanity – in the end, no one’s felt life is better or worse than another’s; we all experience the full range.” Cross-culturally, our measurement approach aimed to ensure the invariance of LoF: if true, it should be true for a villager in rural India as much as for an executive in New York. It’s a bold claim of human universality, which in itself is a culturally sensitive topic (people rightly point out differences in emotional expression and experience across cultures). We tackled this by validating HCI in multiple cultures, and if LoF were to be studied globally, we’d involve researchers from each culture to avoid bias. So, while the idea may land differently – embraced readily in some places, met with skepticism or offense in others – the test of it would be universal.

25.5.2 Avoiding fatalism and passivity

One of the biggest concerns in presenting LoF is that it could be misconstrued as determinism or fatalism. If someone thought “Well, if everything’s going to even out, why bother doing anything? I can’t change the net outcome,” that would be a grave misunderstanding and misapplication. We addressed this repeatedly: LoF is not magic or fate ensuring balance *regardless of your actions* – it works *through* your actions and those of others. If you sit back and let life happen to you, the balance might come in very harsh ways (because you’re not actively opening channels for gentle compensation). For instance, someone depressed who doesn’t seek help might eventually get some relief perhaps in the form of emotional numbness or a breakdown that forces intervention – outcomes far worse than proactively getting therapy or support. Yes, that eventual numbness might count as “some relief” (balance trying to happen) but at unnecessary cost.

There’s also a danger of self-fulfilling prophecy in a twisted sense: if you fear LoF, you might sabotage your own happiness. I’ve heard people say half-jokingly, “I’m afraid to be too happy because life will smack me down afterward.” This is akin to the superstition of not jinxing good fortune. If internalized, that mindset can ruin present joys or lead someone to avoid opportunities (they think, “If I marry this person and become very

happy, maybe something awful will happen to equalize it; better not get too happy”). We must strongly discourage that line of thinking. Life will have ups and downs anyway; refusing happiness is not a shield against future pain, it only ensures you miss out on genuine highs you could have had. The better interpretation is to cherish joys while they last and cultivate resilience for when they ebb, not to preemptively dampen joy.

Similarly, one might misuse LoF to dismiss others’ problems: “Oh, you’re suffering now? Don’t worry, you’ll get over it, things will balance.” That’s callous and could lead to neglect. Or, “This person is extremely happy now; something bad will happen, just wait.” That’s schadenfreude and cynicism. Ethically, adopting LoF should make us *more compassionate*, because we recognize that behind a happy face there were or will be sorrows (so we stay humble and kind, not envious), and behind a sorrowful face there were or will be joys (so we maintain hope and support, not pity them as doomed). It should also instill a sense of responsibility: if a friend is at a low, perhaps you might be one agent through which some balancing high comes (like being there for them).

We explicitly warn in the book that LoF is not an excuse for inaction or injustice. If someone is in poverty or suffering abuse, one must not say “life will sort itself out.” That’s morally wrong. If anything, LoF – by highlighting how suffering demands relief to complete the balance – should spur us to *provide* that relief. Perhaps the universe has many instruments to execute its balancing act, and we are meant to be those instruments (e.g., delivering care, fighting for justice). A law-like tendency doesn’t obviate agency; it works through collective agency. After all, if LoF is natural, it might be realized via natural means – including human intervention. One might poetically say, “If you believe life balances, be the balm for others’ pain and the grounding for others’ joy.” In daily terms: when someone’s extremely down, we help lift them; when someone’s extremely up (and maybe reckless), we gently keep them safe.

25.5.3 Communicating the hypothesis carefully

In a media-driven world, a claim like “Scientists suggest life is exactly fair in the end” would be clickbait and likely misunderstood. As the proposers of LoF, we have a duty to communicate with precision and caution. We’ve emphasized throughout that LoF is a hypothesis on trial, not a proven law. So any public communication must stress *if* it’s true, not that it is true. We must also highlight the caveats: it’s not about morality, it’s not encouraging passivity, it’s not immediate (it’s lifetime-scale, which means it doesn’t stop suffering from happening now, it only says it will be compensated later, which is cold comfort unless we act to make that compensation happen sooner). We also note that it’s not proposing everyone has equal material outcomes or that society is fair – it’s purely about subjective emotional experience.

One scenario to dread is a shallow headline: “Study finds no one is happier than anyone else by death.” This could spark backlash from communities advocating for social justice (“Are you saying the oppressed are as happy as the privileged by the end? That’s outrageous!”). If misinterpreted, it might seem to minimize real inequities. We have to make it clear: LoF is not saying those inequities are okay or don’t matter – it might imply that those who suffer under them will find some personal solace or upswing eventually, but that’s not a justification to allow oppression! If anything, it condemns oppression as a pointless infliction of suffering that must then be alleviated by other means (often requiring enormous efforts, so why do it in the first place?).

Thus, a portion of Chapter 24 and 25 has been almost a “press release to the future,” guiding how to talk about this idea. If evidence accumulates in favor, we plan to *introduce it gradually*, in nuanced forums, and with voices from different fields to contextualize it. The involvement of ethicists, clergy, psychologists, and activists in the conversation is crucial to avoid misapplication. It’s akin to how findings about genetics or IQ must be handled carefully to prevent them from being twisted – LoF has a socio-ethical dimension that demands similar care.

The hopeful side is that if communicated well, LoF could be an inspiring narrative that *unites* rather than divides. It tells us we all travel through the valley and the mountain; empathy is easier when you accept that. It can encourage planning for the long term: governments might invest more in mental health for the elderly, recognizing that trauma earlier in life often needs processing later. It can validate palliative care and hospice philosophies of intense focus on end-of-life peace – not as giving up, but as ensuring that final balance is as positive as possible.

25.5.4 Envisioning a fairness-aware society

Let’s imagine a society that has absorbed the lessons of LoF (whether or not LoF is absolutely true, these would be beneficial moves). In such a society:

- Individuals would be taught emotional literacy and balance from a young age. People would grow up understanding that chasing extreme highs has consequences and that enduring lows will eventually relent. This might encourage a culture of *moderation*, resilience, and patience. Someone facing a setback might think, “This is awful now, but it’s not the end of my story. I can work through it, and there will be better days.” Someone riding high might practice gratitude and not arrogance, perhaps thinking, “I’ve been fortunate; I should use this good period well and be prepared to handle future challenges.”

- Therapists and caregivers would integrate this perspective by instilling hope in patients that no depression or grief is final. Therapy might explicitly work on finding or creating balancing experiences: for trauma, finding restorative experiences; for mania, grounding techniques. End-of-life care, as mentioned, would prioritize emotional closure. There might be a commonplace practice of “life reconciliation” sessions for elders – structured opportunities to reflect, forgive, and make peace, akin to what hospice chaplains do but available more broadly.
- Communities would perhaps develop norms of stepping in when someone is extremely down (seeing it almost as a natural duty: “they are in a deep trough, we must help raise them”) and likewise gently tempering extremes (like preventing dangerous overindulgence or hubris). This doesn’t mean punishing happiness; rather, a community might encourage those who have abundance of joy or resources to *share* – effectively distributing happiness (think of a feast: one person’s great fortune becomes everyone’s celebration).
- Policies might aim at what one could call “experiential equity.” For instance, social safety nets would ensure that people hit by hard times aren’t left without recourse – because we know if they remain in suffering too long, the eventual “compensation” might come too late or in unhealthy ways. There could be greater emphasis on rehabilitation in justice systems: not simply inflicting more pain on criminals, but trying to ultimately reintegrate them so their story can move toward neutral rather than end in suffering (which, under LoF, doesn’t serve anyone – one could even cynically argue that if a criminal is left to suffer horribly in prison until death, LoF would imply they must have had a lot of unsanctioned joy earlier or something; better to balance them through remorse, restitution, and then some form of peace). Educational systems might incorporate life-balancing skills: coping strategies, mindfulness, community service (to connect disparate life experiences).
- Justice in a fairness-aware society would focus on healing. Retributive justice (an eye for an eye) is actually a crude human attempt to impose balance – you made someone suffer, you should suffer. LoF frames balance in terms of experience, not moral ledger, but a humane justice system could say: the goal is to restore victims and rehabilitate offenders so that net suffering is minimized going forward. Punishment might be used only as necessary to correct and deter, not to add suffering for its own sake (because adding more suffering doesn’t create good – it just adds more negative ledger that will need balancing somehow, maybe through

state guilt or cycles of crime, etc.). Instead, a restorative justice model – making amends, reconciliation – aligns with the idea of closing the loop positively.

- Economics and work might also be influenced. Perhaps there'd be more emphasis on work-life balance (since chasing extreme success at work at the cost of personal life might lead to an eventual emotional crash). Social security systems could treat unemployment or retirement not just as economic issues but as emotional transitions that need support (to avoid identity crises and despair that could tip a life's balance negatively at the end).
- Media and culture might reflect more nuanced stories: rather than idolizing those who “have it all” or pitying those who have nothing, narratives might focus on the arc of life – the comeback stories, the hidden struggles behind smiles. We might celebrate empathy and kindness as highly as we celebrate winning and fame, seeing those as mechanisms by which society evens out fortunes ethically.

This might sound utopian, but it's essentially extrapolating existing positive trends. Many of these ideas are already present in progressive thought – LoF would bolster them with a unifying rationale: to ensure no one's life gets *too* far out of balance where possible, and when it does, to intervene timely.

We described in Chapter 24 how practical initiatives could align with LoF: e.g., “widen channels” meaning always make sure help is accessible (so that negative experiences can be countered), reversibility in design meaning let people undo mistakes (so one bad choice doesn't doom their trajectory – think of forgiving debt, second-chance programs, etc.), and finish repairs meaning don't leave people in permanent limbo of pain if you can help it.

25.5.5 Ethos of balance and compassion

At a societal ethos level, LoF promotes two key values: long-term perspective and compassion. Long-term perspective means we encourage thinking beyond immediate gratification or panic. Someone in a crisis might remember that life is long and capable of change – reducing impulsive harmful acts like suicide or violence (given time, things can shift). It also means policies are evaluated on how they affect a person's whole life, not just the next quarter or election cycle.

Compassion is reinforced by the idea that we all partake in the human condition equally at the end. It invites a kind of radical empathy: the millionaire and the beggar, at life's close, both will have known deep sorrow and joy; thus, we should treat both with humanity now. It discourages arrogance during good times and disdain for those in bad times.

In Chapter 24, we phrased a personal takeaway as a credo: “Every hurt calls for a balm, and every joy carries a duty of care.” This neatly summarizes the societal ethic from LoF. Every instance of suffering we encounter (in ourselves or others) is a cue to provide some healing or support – to supply the compensating positive because that’s how balance is achieved. And every instance of joy or privilege we have is a cue to be responsible with it – not to hoard it or abuse it, but perhaps to use it to prevent future pain (our own or others’). Imagine if individuals and governments acted on that: rejoicing in prosperity, but immediately thinking, “how can we use this prosperity wisely to alleviate suffering or secure well-being for when harder times come?” and responding to every disaster or hardship with swift aid, knowing that “this must be countered, we can’t leave people in the red.”

In practice, we see glimpses of this: charitable giving spikes after tragedies (balance impulse), and wise leaders invest surplus in safety nets (preparing for future lows). LoF would encourage making that the norm rather than a special effort.

Summarizing, a fairness-aware society wouldn’t be passive waiting for cosmic balance – it would be actively balancing as a matter of policy and principle. Perhaps ironically, even if LoF eventually turns out not to be a strict law, living as if it *should* be true might create a kinder, more resilient society. In that sense, it’s akin to a guiding ideal (like how we hold equality as an ideal even when reality falls short).

We must also accept that even in such a society, tragedies will happen. LoF or not, some lives do end tragically imbalanced currently. Our responsibility then is to learn from each case – to find what prevented balancing (was it isolation? stigma? lack of resources?) – and then fix those gaps for the future. That way, even the exceptions push us to improve conditions such that they become rarer.

25.5.6 Where we go next:

We’ll step back from these detailed analyses and implications to reflect on the journey we’ve taken, the knowledge we’ve gained, and the road ahead. It’s time to confront how this idea stands before the scrutiny of the world’s wisdom and evidence, and to offer our final words on what it means personally and intellectually. Let us move to the final reflections, where we will address the ultimate question: *Where do we stand with the Law of Fairness at the end of this long exploration, and what do we do with it?*

25.6 Final Reflections

As we conclude this extraordinary journey through the Law of Fairness, we step back to assess what we have accomplished and what it means. Writing this book has been an exercise in bridge-building – bridging science and spirituality, data and hope, rigor and compassion. Now, in these final reflections, we want to consider how prepared we are to present this idea to the wider world and to its most serious critics, and to acknowledge the uncertainty and excitement that come with reaching the end of such an exploration.

25.6.1 A bold endeavor

This work was an ambitious interdisciplinary endeavor from the start. Few ideas dare to propose a new law of nature that intersects with something as deeply personal as human joy and suffering. We did so fully aware that it would invite skepticism from all quarters – and indeed we *welcome* that skepticism, because only through stringent testing will we know if this idea holds water. The hypothesis is unquestionably bold: it asserts a strict symmetry in subjective experience, a domain known for its complexity and variability. Such boldness was necessary to make the question concrete. By staking a claim as absolute as LoF, we made it possible for evidence to clearly support or refute it; a fuzzier claim would languish in ambiguity. So we stand by the boldness, with the important caveat that the final verdict rests with evidence. We have no intention of dogmatically clinging to LoF if data and reason show it to be false. Life may turn out to be only approximately fair, or fair only under certain supportive conditions (like having community, or not experiencing extreme trauma). Those outcomes would themselves be incredibly valuable to know. Imagine if research finds that life can be fair, but only when society provides specific supports or when a person has developed certain resilience skills – that would direct us to improve those conditions for everyone. Or if we find life is, say, 80% fair – most people come close to balance but some don't – that still deepens our understanding of well-being and its limits. In science, even a disproven hypothesis can illuminate truths indirectly, by sharpening the questions and methods.

25.6.2 Shifting the dialogue

Regardless of LoF's ultimate fate, this project has already achieved something important: it has reframed the question of life's fairness from a matter of opinion or faith into a matter of inquiry. Traditionally, asking "Is life fair?" was almost a rhetorical or philosophical question, often leading to clichéd answers or resignation ("life isn't fair, deal with it"). We turned it into a specific empirical hypothesis: "Do the integrated experiences of a lifetime sum to zero?" and "What evidence would demonstrate or falsify that?" By doing so, we invited a new kind of dialogue. It's no longer about trading anecdotes ("I knew someone who suffered all their life... well, I knew someone who had

it easy...") but about gathering data in a systematic way across many lives and contexts. That is a significant shift. We've moved the conversation into a space where it can involve psychologists, neuroscientists, statisticians, ethicists, theologians – all looking at the *same* defined question, even if from different angles. In bridging disciplines, we had to create a common language (defining hedonic units, neutral levels, etc.), and that itself is progress. Even if the answer turns out to be "no, life isn't strictly fair in this sense," we will have learned a great deal in the process of checking. We will have developed methods to measure life outcomes, perhaps discovered partial balances or key exceptions, and overall made the discussion more nuanced and evidence-based.

25.6.3 Responsibility and empowerment

One outcome of treating fairness as a testable hypothesis is that it compels us to take responsibility for it. Instead of shrugging and saying "sometimes life sucks, nothing to do about it," we are motivated to study it, and by extension, to *do* something about it. If indeed suffering tends to be balanced by relief given the chance, then our job as a society is to maximize those chances (through policies, interventions, mutual aid). If we suspect LoF might be true, we don't just wait for balance to magically happen – we actively facilitate it (because our actions may be the very mechanism by which it happens). Conversely, if we find evidence that some lives end with serious imbalances, that places a moral onus on us – we can't assume any hidden hand will fix those, so it's up to us to create fairness through deliberate action (justice, therapy, support). In either case – whether LoF holds or fails – the approach of examining it scientifically *enlightens our sense of responsibility*. We can't hide behind platitudes or throw up our hands. We're forced to look at suffering and joy in the clear light of data and say, "What patterns do we see? What can we do with this knowledge?" The process of research here is not passive observation; it inherently suggests interventions (because if you find, say, a factor that prevents balancing, you'd want to address that).

25.6.4 The ultimate measure – data

At the end of the day, the truth of LoF will be decided by evidence. All the beautiful analogies, philosophical musings, and theoretical models we've discussed must bow to what reality shows us. Throughout the book, we have tried to marry hope with empiricism: there is a hope that life has this kind of fairness, but hope must be put to the test. We've outlined what those tests look like: long-term longitudinal studies with robust affect measurement, cross-cultural lifespan analyses, etc. These are ambitious, costly studies that might take decades – which means patience and perseverance are required. We might not have a definitive answer in our lifetimes, and we should be at peace with that. Science is often a relay race through generations.

Until the data can speak clearly, LoF remains a well-defined question mark. And that's okay. In fact, it's more than okay – it's exciting. We took an amorphous question ("Is life fair?") and turned it into a concrete research program. That alone is an accomplishment in the realm of ideas. Now it lives in the hands of the scientific community and, in a way, in the hands of every person who might reflect on their own life or contribute their story to a dataset. The ultimate measure of this theory is literally in measures: those cumulative HCl curves, those distributions of end-states, those deep interviews with people at life's end. We have tools and paths to get those answers. It may take time and new technology (maybe lifelogging or large-scale data from wearables combined with psychological assessments could accelerate this – something we hint at in Part V). However long it takes, we can find deep meaning that the journey itself – the pursuit of LoF – is yielding insights. We're learning how to measure well-being better, how different life events impact trajectory, how adaptation works over years... these are rich contributions irrespective of LoF's final status.

25.6.5 No platitudes, only hypotheses

From the outset, we made a deliberate choice to avoid comforting platitudes. Now, at the conclusion, we reiterate that stance. This book does not say "Don't worry, be happy, everything happens for a reason." It doesn't guarantee a happy ending in the simplistic sense, nor does it justify suffering as deserved. Instead, it presents a hypothesis with teeth – one that can bite back if false. We should strive to handle nuance and uncertainty, rather than seeking comfort in a warm blanket of unfalsifiable reassurance. We tried to model the very principle of fairness in our inquiry: we looked at hard truths (children suffering, random tragedies, etc.) without flinching, and we also entertained hopeful possibilities without sneering. This balanced approach – optimistic but critical – is perhaps one of the best outcomes of pursuing LoF. It trained us to be fair-minded in our evaluation of evidence and theories. As written earlier, to truly test if reality is fair, we must approach reality fairly: without bias, without despair or Pollyannaism, willing to see whatever is there.

In a sense, the process of investigating LoF required a kind of intellectual virtue that hopefully comes through in the writing. We had to hold in our heads both the worst and best of life and not let either alone dominate our judgement. It's easy to become cynical looking at the world, and easy to become naively hopeful by ignoring problems – it's much harder to stare at tragedy and still say, "I hypothesize there might be a hidden symmetry that doesn't erase the tragedy but puts it in a larger context that also contains redemption." That's a nuanced position. Whether or not nature vindicates it, adopting

that stance – *neither blind to pain nor to hope* – is something to carry forward as a way of living and thinking.

25.6.6 A more nuanced conversation

Already, we see that discussing LoF has made people’s conversations about life’s ups and downs more nuanced. Those who have engaged with our drafts or talks didn’t just say “agree” or “nonsense” – they started asking interesting questions: “What about people with disabilities? What about extreme cases like Holocaust survivors? How do we count that?” or “If it’s true, could it relate to the brain’s predictive mechanisms or maybe to social support dynamics?” These are much richer dialogues than the fatalistic shrug or the simplistic optimism that usually ends such conversations.

The conversation is also more empowering. It’s not just “life is unfair, period” or “maybe God will reward in heaven” – it’s “what patterns and mechanisms might make it fair or unfair, and how does that inform our actions now?” When anchored to empirical pursuit, even a philosophical question gets invigoratingly concrete. It also humanizes science: talking about balancing sorrow and joy over a lifetime is not dry or abstract; it’s something every person can relate to, yet we’re tackling it with science’s tools. It can elevate both the science and the human story – making science more compassionate and our existential discussions more grounded.

25.6.7 Recruiting the best thinkers

The measure of success for this book is not that everyone agrees with it, but that it sparks rigorous engagement across disciplines. Ideally, it will inspire collaborative efforts: maybe a neuroscientist reaches out to help refine the HCl with brain data, or a sociologist wants to test it in different cultures. The “world’s best thinkers” engaging with LoF – whether to dismantle or build on it – is exactly what we need to get to the truth.

In that sense, we should consider this book a conversation starter, not the final word. We confronted many angles here internally, but the broader intellectual community will bring fresh perspectives. We have tried to leave no obvious question unexamined – so if new ones arise, they will likely be interesting, non-obvious ones that push the idea into new territory (e.g., maybe someone connects it to evolutionary biology in a new way, or to information theory in AI systems, etc.). That’s exciting. Whatever happens, LoF will have served to galvanize thought and research that wouldn’t have happened otherwise.

25.6.8 Carrying the inquiry forward

As we close this book, consider carrying this inquiry forward in your own way. This is not a neat, resolved ending; it’s the end of a chapter and the beginning of a collective

investigation. You might test LoF in the laboratory of your own life: reflecting on your experiences, perhaps you'll keep a "balance diary" and notice the ebbs and flows. Or you might discuss it with others, trading perspectives and stories, which itself can be enlightening. If you're in a position to formally research it, by all means join the effort: gather data, propose alternative theories (maybe there's a "Law of Proportionate Fairness" or a threshold effect or something we missed). If you feel skeptical, design the critical test— and then do it. If you feel hopeful about it, think of ways to bolster the case empirically or to apply its principles to help people.

Above all, remain critical yet open, and compassionate. Those are the twin virtues we've tried to cultivate in these pages. Critical and open is how science progresses (neither cynical dismissal nor gullible acceptance, but a fair weighing of evidence). Compassionate is how society progresses (treating each other with understanding that everyone's highs and lows are part of being human).

In the end, whether the Law of Fairness stands as a new scientific principle or falls as an overreach, it already shone a spotlight on what truly matters: reducing suffering and nurturing joy. If LoF is true, it reveals a strict structural constraint on experience – that every tear and laugh are balanced in the integrated ledger by the death of mind. If it falls, it underscores how much work we have to do to create fairness where nature doesn't. In either case, the message is clear: our choices matter. If life is fair by itself, our choices determine *how* that fairness plays out (whether through kindness or cruelty). If life isn't fair by itself, our choices determine *whether fairness exists at all*.

And perhaps that realization is the final synthesis: that a fair life – in terms of meaning and care – is something we co-create, law or no law. The ledger of life is not just a passive tally but an account we can actively manage through love, effort, and knowledge. As we go forth, let's continue to seek the truth of our experiences with both boldness and humility. And let's continue, above all, to be kind to one another, for in kindness we find a form of fairness we can guarantee.

Summary of Main Ideas

Part I — The Question That Won’t Go Away

Chapter 1 — Why Fairness? Why Now?

Promise: Introduces the core question of whether life can be literally fair in felt experience, and why this question demands attention today.

Outcomes:

- Grasp through real-life stories (a child, a rockstar, a monk) why the fairness of life’s outcomes is a pressing, personal question.
- Differentiate procedural, distributive, predictive, and experiential fairness, and see why only the experiential notion can adjudicate a life’s actual felt outcome; LoF addresses experiential fairness exclusively and makes no moral or metaphysical promises.
- Understand the roadmap of the book: definitions first, then measurement, then tests and fail criteria, all preregistered and audited; the work is non-teleological and non-moralizing, treating fairness as a descriptive constraint to be confirmed or rejected by data.

Subsections:

- 1.1 A Child, a Rockstar, a Monk
- 1.2 What People Usually Mean by Fairness
- 1.3 The Hard Cases No One Likes
- 1.4 Why “Tendency Toward Fairness” Isn’t Enough
- 1.5 How This Book Works

Chapter 2 — Feelings as the Final Currency

Promise: Makes the case that subjective feelings (pleasure and pain) are the fundamental currency of fairness, more important than any external marker, and explores whether and how such feelings can be rigorously measured.

Outcomes:

- Understand why resources or achievements only matter via how they *feel* to the person living them, establishing feelings as the ultimate metric of a life's fairness.
- See how feelings can be quantified responsibly using a hedonic composite built from first differences across channels (self-report, autonomic physiology, brain signals, behavior, dreams), with reliability weighting, cross-validation, and measurement-invariance checks so that scores are comparable across persons, cultures, and states.
- Get a one-page tour of affective science that anchors measurement in biology: core affect (valence and arousal), nociceptive and interoceptive circuits (insula, anterior cingulate), valuation and control systems (vmPFC, rIFG), and neuromodulatory influences (dopamine, serotonin, noradrenaline) — framing $HCI\Delta$'s channels in mechanistic terms.
- Learn the ethical nonnegotiables for studying feelings: relief is a systems variable (never withheld), participants' comfort and dignity override data capture, and protocols are designed so the ledger estimate $\bar{L}(t)$ never becomes a pretext for allowing harm.
- Preview the key measurement concepts: Hedonic Composite Units (HCU) obtained by integrating $HCI\Delta$ over time and calibrated for invariance; a cumulative life ledger $\bar{L}(t)$ defined as the time-integral of $HCI\Delta$; and a neutral band K at closure for equivalence testing in Part II.

Subsections:

- 2.1 Why Feelings, Not Just Things
- 2.2 Can Feelings Be Counted?
- 2.3 Somatic Markers
- 2.4 Affective Neuroscience in One Page
- 2.5 What We Will Never Do
- 2.6 Preview: Hedonic Composite Units (HCU) and the Life Ledger

Part II — The Law, Stated Clearly

Chapter 3 — The Law of Fairness

Promise: States the Law of Fairness plainly and formally, so there's no ambiguity about what is claimed, and lays out the logical structure that makes it testable.

Outcomes:

- Read the canonical statement in one sentence, with symbols defined: for a unified conscious stream from onset to death of mind T , $L(T) = \int_0^T F(t) dt = 0$, with identity, pause, and split/merge rules stated up front and a measured proxy $\hat{L}(T)$ built from $HCI(t)$ (with $HCI\Delta$ as the discrete first-difference operationalization) for empirical tests.
- Understand six core assertions: (1) ledger closure at death of mind ($L(T) = 0$; evaluated operationally within preregistered $\pm\varepsilon$ bounds); (2) one ledger per unified stream with preregistered split/merge/pause rules; (3) the law is a passive constraint, not a purpose; (4) adaptation and baseline trends are insufficient to imply closure; (5) dreams and narrative counterweights can contribute measurable Δ toward neutrality; and (6) a clear distinction between the latent ledger $L(T)$ and its empirical estimate $\hat{L}(T)$ with explicit error budgets.
- Learn what the law explicitly does not claim (no guarantee of equal happiness for all, not a moral reward system, not about population averages) to avoid common misinterpretations.
- Identify the boundary conditions where the law applies or fails (e.g. how the ledger pauses in dreamless sleep or anesthesia, how multiple personalities or split brains are handled, how very short lives or non-human minds might be treated).
- Grasp the definition of “death of mind” as the endpoint for evaluating the ledger, and why the law is checked at the irreversible end of consciousness rather than biological death alone.
- (Research Notes outcome): See the formal ledger integral $L(T) = \int_0^T F(t) dt$ expressed in HCU and the mathematical criteria for “neutrality” (equivalence bounds, etc.), linking the intuitive claim to a precise equation that can go into a preregistration.

Subsections:

- 3.1 Canonical Statement
- 3.2 Six Assertions of the Law
- 3.3 What the Law Does Not Say
- 3.4 Boundary Conditions
- 3.5 The Death of Mind
- 3.6 Research Notes: The Ledger Integral and State–Change Formalism

Chapter 4 — Constraint, Not Purpose

Promise: Explains the nature of the Law of Fairness as a passive constraint (a “guardrail”) in how lives unfold, as opposed to any guiding purpose or cosmic intent, and places it in the context of scientific laws versus teleological ideas.

Outcomes:

- Understand the crucial difference between a constraint and a teleological force: why the Law of Fairness, if true, is like a rule limiting possible outcomes (e.g. like energy conservation), not a goal that the universe or individuals are actively pursuing.
- Learn why viewing the law as a constraint avoids mystical thinking and “just-so” optimism — it doesn’t mean lives are *guided toward* balance, only that non-balancing paths get naturally curtailed by the system’s dynamics.
- See the argument for why “constraints beat miracles”: framing a fair-outcome phenomenon as emerging from many small regularities (micro-level processes like adaptation, feedback loops, etc.) is scientifically stronger than invoking a grand purpose or cosmic justice.
- Situate the Law of Fairness within philosophical accounts of laws: understand the idea of a best-system law (a concise summary of regularities) versus a governing law, and see which camp this claim would fall into and why that matters for testing it.
- Recognize explicit language choices we make to avoid teleology: for instance, always describing processes in terms of feasibility and selection (Queue System mechanics) rather than saying “nature wants X.” This outcome highlights how we communicate the law to keep it in the scientific lane.
- Appreciate how avoiding teleology still allows meaningful predictions and tests: even without attributing purpose, the law would produce distinctive empirical signatures (outlined in Ch. 3 and tested later), which is what makes it a candidate for real scientific knowledge rather than a philosophical musing.

Subsections:

- 4.1 Guardrails vs. Steering
- 4.2 Why Constraints Beat Miracles
- 4.3 No Teleology
- 4.4 Lawhood: Best System vs. Governing Law
- 4.5 Research Notes: Optional Stopping and Regularity

Part III — How the System Works (From the Inside)

Chapter 5 — The Queue System (QS)

Promise: Introduces the hypothesized mechanism that enforces the Law of Fairness constraint: the Queue System (QS). QS posits that an individual's available choices and mental policies are pruned/limited so that only trajectories compatible with terminal neutrality remain viable over time.

Outcomes:

- Comprehend QS in a sentence: “a regulatory process that limits which actions or thoughts ‘line up’ next, selecting admissible next steps that keep compensation feasible.”
- Understand choice sets and admissible policies: at any given moment, not all imaginable actions are available — QS ensures that only paths that would not force an irredeemably imbalanced ledger remain viable in the long run.
- Learn about possible neural correlates of QS: rIFG (inhibitory control), ACC (conflict monitoring), vmPFC (valuation), and insula (interoception) as candidate hubs.
- Discover a novel interpretation of dreams: dreams as “low-cost counterweights” that can contribute to the ledger without high real-world cost.
- Know what would falsify QS: after controlling for known factors, absence of measurable choice-set restriction following large pains/pleasures, or clear cases where obviously harmful options were freely chosen and led to irreparable imbalance without any QS-like braking.

Subsections:

- 5.1 QS in a Sentence
- 5.2 Choice Sets and Admissible Policies
- 5.3 Neural Correlates: rIFG, ACC, vmPFC, Insula
- 5.4 Dreams as Low-Cost Counterweights
- 5.5 Research Notes: QS-Residuals After Nuisance Modeling
- 5.6 What Would Falsify QS?

Chapter 6 — Time Horizons and the Shadow Price

Promise: Explains why balancing pressure rises as the expected remaining horizon H_t shrinks and introduces the shadow price λ_t —a horizon-dependent multiplier on expected terminal imbalance that typically grows as opportunities to compensate dwindle. Specifies observable endgame signatures and explicit falsifiers.

Outcomes:

- Why the endgame sharpens choices: With a binding terminal neutrality constraint, admissible trajectories narrow as H_t contracts; λ_t rises, producing “last-window” behavior (reconciliation, closure, relief-seeking).
- Formal idea of λ_t and H_t : Define $H_t = E[T - t]$. With $L = F(t)$ and convex terminal loss $\phi(L_T)$, the costate $\lambda_t = \partial J / \partial L$ (where J is the value function) acts as the shadow price on terminal imbalance and generally increases as H_t shrinks, raising the weight on closure-improving actions.
- Not discounting but constraint-induced urgency: λ_t reflects a terminal boundary condition, not present bias; it predicts targeted balancing moves rather than indiscriminate “now-favoring.”
- What to measure: dACC/IFG urgency signals; pupil/HRV effort indices; time-perspective shifts (positivity/meaning)—all should scale with perceived horizon length.
- Population-scale horizons: When cohorts share endpoints (e.g., fixed institutional deadlines, retirement windows, crisis end-dates), a population-level shadow price may emerge, yielding policy windows where small interventions create outsized fairness gains (e.g., facilitated reconciliation or relief access).
- Fail patterns: No endgame intensification, no rise in urgency/effort markers, or stable/expanding cross-person HCl variance near the end—if generalized, this falsifies the mechanism.

Subsections:

- 6.1 Why Endgame Balancing Intensifies
- 6.2 The Intuition of Shrinking Horizons
- 6.3 Hospice Across Cultures
- 6.4 Research Notes: Shadow Price λ and Horizon H
- 6.5 What to Measure (EEG/fMRI, Time Perspective)
- 6.6 Fail Patterns for Horizon Scaling
- 6.7 Population Shadow Price and Policy Windows

Part IV — Measuring Feeling Without Fooling Ourselves

Chapter 7 — The Hedonic Composite Index (HCI)

Promise: Introduces HCI as a multi-channel, uncertainty-aware measure of momentary felt experience and defines the Hedonic Composite Unit (HCU) used to total experience over time. Establish safeguards so measurement can guide, not game, science.

Outcomes:

- Five inputs, one signal: HCI fuses self-report, physiology, brain signals, behavior, and dream content into a single latent estimate with calibrated uncertainty, outperforming any single channel when channels disagree or drift.
- Why composite beats single meters: Aggregating partially independent indicators reduces bias (e.g., social display in self-report) and exposes contradictions that force better models instead of comforting stories.
- Keeping it honest: Preregistration, randomization/blinds where feasible, and third-party audits prevent cherry-picking; privacy and consent are built in so participation remains humane.
- Under the hood (latent + state-space): Confirmatory factor/IRT models define channel loadings; a state-space model tracks HCI through time, yielding smooth estimates with explicit uncertainty bands.
- From HCI to HCU and ledgers: Integrating HCI over time produces HCU and a running ledger $\bar{L}(t)$; we explain how uncertainty propagates so lifetime claims carry intervals, not bravado.
- Known fail conditions: Systematic channel contradiction, adversarial “gaming,” or non-stationary drift that breaks calibration are treated as red flags that pause inference and trigger redesign.

Subsections:

- 7.1 Inputs: Report, Physiology, Brain, Behavior, Dreams
- 7.2 Why Composite Beats Single Meters
- 7.3 Keeping It Honest: Blinds and Preregistration
- 7.4 Research Notes: Latent Models (CFA/IRT) and State-Space
- 7.5 Hedonic Composite Units (HCU)
- 7.6 Fail Conditions for HCI

Chapter 8 — “Same Scale” Across People and Places

Promise: Solves the comparability problem: when we say two ledgers are equal, do they mean the same thing across individuals or cultures? Formalizes measurement invariance and shows how to anchor, calibrate, and carry uncertainty into the ledger.

Outcomes:

- Define the invariance ladder: We require configural → metric (→ scalar where powered); if metric fails, cross-group comparisons stop, and claims are restricted to within-person change.
- Culture and age effects handled directly: Norms, language, and cohort baselines can shift reporting or physiology; the chapter shows how to model and test these influences rather than averaging them away.
- Universal anchors: Pain (ice bath), chills from inspiring music, and social exclusion provide shared anchors; careful translation and safety checks let us map “+1 HCl” in one group to “+1 HCl” in another.
- Calibration ladder: First within-person calibration, then between-person using overlap tasks, then cross-cultural extension with multiple anchors; each rung supplies evidence, or limits the scope, of fair comparison.
- Propagating uncertainty: We report ledger estimates with intervals (e.g., “0 ± x HCU at 95%”), not point claims, so neutrality at life’s end is a statistical statement that can be confirmed or refuted.
- Fail and narrow: If only configural or metric holds, we narrow claims; accordingly, if scalar holds, we allow absolute cross-group comparisons but still report the uncertainty we carry forward.

Subsections:

- 8.1 The Invariance Problem
- 8.2 Culture and Age Effects
- 8.3 Universal Anchors (Pain, Chills, Social Exclusion)
- 8.4 Research Notes: Configural → Metric → Scalar Invariance
- 8.5 Calibration Ladder (Within → Between → Cross-Cultural)
- 8.6 Propagating Uncertainty into the Ledger

Part V — Identity and Edge Cases

Chapter 9 — Unity of the Stream

Promise: Defines who has a ledger by specifying when multiple processes constitute one unified stream and how we adjudicate pauses, splits, and edge cases. Establishes principled rules for identity without teleology or speculation.

Outcomes:

- Why fairness needs a subject: Under LoF the system must close the ledger for each unified stream; we operationalize “unified” via integrated information and control—cross-influence, shared memory, and coordinated action.
- The Unity Index (plain speech): A pragmatic score from observable signs (e.g., synchronized brain activity across regions, cross-talk between subsystems, a coherent self-model) helps decide “one stream vs. two.”
- Pauses vs. closures: Sleep, anesthesia, and short reversible comas count as pauses of the same ledger; irreversible loss of integration marks its end.
- Splits and merges: Split-brain/DID are handled by whether there is ongoing cross-talk; separate ledgers apply when independent access and control persist, with explicit rules for later merging if integration returns.
- Non-traditional minds: Criteria extend to AI or organoids by requiring the same observable integration before granting a ledger; adjudication is blinded where possible to avoid motivated reasoning.
- Ethics and Fail patterns: Identity calls affect who is counted; we err conservative, acknowledge uncertainty, and state thresholds. “Relief is a systems variable; comfort and dignity override data collection.”

Subsections:

- 9.1 Conscious Access: One Stream or Two
- 9.2 The Unity Index (Plain Speech)
- 9.3 Pauses: Sleep, Anesthesia, Coma
- 9.4 Split-Brain and DID
- 9.5 AI and Brain Organoids
- 9.6 Research Notes: Blinded Adjudication and Thresholds

Part VI — Evidence We Can Look For (Right Now)

Chapter 10 — Dreams: The Night Workshop

Promise: Proposes that dreams help counterbalance unresolved daytime loads at low real-world cost and spells out measurable predictions. Treats dreams as testable signals about ledger repair, not as mystical messages.

Outcomes:

- What dreams could do for the ledger: Under LoF the system must find low-cost counterweights; REM dreams are hypothesized to simulate and resolve open affective “tickets,” shifting the next day’s HCl toward neutral.
- Classic observations reframed: Recurring themes, rebound after suppression, and emotionally charged dreams are presented as candidate signatures of balancing—not proof—ready for preregistered tests.
- Predictions to test: After unusually negative days, dream affect should tilt positive (valence inversion), and REM intensity/duration should scale with the size of the day’s imbalance; the reverse holds after unusually positive days.
- How to measure well: We outline REM timing, sampling strategies, and blinded content coding to avoid bias (e.g., alarms, journals, independent raters), with uncertainty reported alongside effect sizes.
- A one-week exercise: Readers can run a classroom or personal mini-study that pairs daily moods with dream valence to see the inversion pattern and learn how preregistered analyses work.
- Ethics and Fail patterns: “Relief is a systems variable; comfort and dignity override data collection.” If dream content simply mirrors daytime mood, or REM deprivation has no long-term effect on balance, the dream mechanism is weakened.

Subsections:

- 10.1 What Dreams Do for the Ledger
- 10.2 Classic Observations
- 10.3 Predictions: Valence Inversion After Tough Days
- 10.4 Research Notes: REM Timing, Sampling, Coding
- 10.5 A One-Week Dream Ledger Exercise
- 10.6 Fail Patterns in Dream Data

Chapter 11 — End-of-Life: Where the Law Shows Its Hand

Promise: Argues that the death of mind is the sharpest test of LoF: if the law holds, a lifetime ledger must close to neutral as terminal consciousness approaches. Specifies exact equivalence margins and uncompromising ethics.

Outcomes:

- Three quantitative gates: We test terminal neutrality/compression with preregistered bounds—mean within ± 0.15 z, slope within ± 0.05 z/day, variance ratio ≤ 0.80 vs. matched baseline—and treat failures as evidence against LoF.
- Convergence from aging psychology. As perceived time-horizons shrink, people preferentially encode and retrieve positive over negative material and report better emotion control. LoF interprets this as $F(t) \rightarrow 0$ dynamics: we predict both mean and variance of HCl(t) compress as the horizon shortens. Design note: align time to death-of-mind, preregister the compression metric, and show that the effect remains after analgesia/sedation windows are treated as exogenous.
- What clinicians report vs. what we measure: Qualitative accounts of reconciliation, peace, or surges of clarity motivate measurable proxies (HCl trajectories, variance shrinkage, neural markers) rather than stand in for them.
- Ethics at the bedside: “Relief is a systems variable; comfort and dignity override data collection.” Only non-intrusive, consented observation (or surrogate consent) is admissible; research must follow care, never lead it.
- Reading anecdotes like data: We separate cultural scripts from signals, use time-stamped logs instead of memory, and guard against cherry-picking by preregistering how we will quantify closure.
- Confounders and counter-examples: We plan for analgesics, delirium, and communication barriers; the chapter names patterns that would contradict LoF (e.g., persistent negative drift with no compression despite comfort).
- Why this test is decisive: If careful studies repeatedly cross all three gates, LoF gains credibility; if they fail cleanly, the law is in serious doubt.

Subsections:

- 11.1 Why This Is the Sharpest Test
- 11.2 What Hospice Workers See
- 11.3 Ethics: What We Will and Will Not Do
- 11.4 Research Notes: Variance Compression and Neural Signatures
- 11.5 Reading Anecdotes: Scripts vs. Signals
- 11.6 Fail Patterns at Terminal Closure

Chapter 12 — The Lab Bench: Horizon Tasks and TMS

Promise: Describes short-horizon experiments that create “mini endgames” and uses noninvasive brain stimulation to probe candidate control hubs. Tells readers exactly what success and failure look like in a lab setting.

Outcomes:

- Short vs. long horizons in tasks: When participants believe time/opportunities are limited, LoF-consistent compensation should intensify (e.g., avoiding “ending on a bad note”); with long horizons it should relax—differences we can quantify within subjects.
- Expected control-hub signatures: Regions like rIFG (inhibitory control), ACC (conflict monitoring), vmPFC (valuation), and insula (interoception) are hypothesized to show distinctive patterns when horizons shrink, consistent with Queue System control.
- Perturbation tests: If we inhibit candidate hubs with TMS, compensatory choices should weaken or invert; sham stimulation and out-of-network targets serve as negative controls.
- Rigor by design: Preregistration, power analyses, and justified ROIs keep studies credible; outcome measures and analysis plans are locked before data are seen to prevent p-hacking.
- Social channel (contagion). Others’ expressed affect can measurably shift our own. We include a social-exposure Δ —e.g., density/sentiment of close-tie messages, in-person interaction diaries, or group mood measures—entered as standardized first differences into $HCI\Delta$. Prediction: higher exposure to positive (negative) social affect yields short-lag increases (decreases) in $HCI(t)$ after adjusting for baseline ties and prior mood; persistent wrong-sign effects are a local falsifier.
- Clear negatives: No behavioral change under short horizons, TMS with no predicted effect, or neural activity inconsistent with horizon engagement would count against the proposed mechanism.
- How this bears on LoF: Lab results don’t prove the lifetime law by themselves, but consistent horizon effects and hub sensitivity would support a concrete path by which a balance constraint could be implemented.

Subsections:

- 12.1 Short vs. Long Horizons in the Lab
- 12.2 Expected Control-Hub Signatures
- 12.3 Perturbation: TMS to rIFG/ACC
- 12.4 Research Notes: Preregistration, Power, ROIs
- 12.5 Negative Controls
- 12.6 Fail Patterns in Lab Tests

Chapter 13 — The Long View: Telemetry Across Years

Promise: Shows how to observe LoF in real time across months, years, or a life course using longitudinal HCI telemetry. Specifies the balancing patterns we should and should not see if the lifetime ledger must close near death of mind.

Outcomes:

- What longitudinal HCI looks like in practice: Practical sampling with phones and wearables, privacy-respecting prompts, and aggregation that yields an individual's ledger curve rather than sporadic mood snapshots.
- Predicted long-term signatures: Under LoF the system must exhibit convergence toward neutral near end-of-life and compensatory swings after large shocks; we preview how to detect both trends and residuals.
- Major events as “ledger shocks”: Distinguish compensation from regression-to-the-mean by preregistered windows, controls, and counterfactuals, so big negatives followed by relief (or big positives followed by challenges) are not misread.
- Methods for gaps and dropout: Use mixed-effects and Bayesian hierarchical models with principled handling of missingness so the observed compression isn't an artifact of attrition.
- Citizen Science without the creepiness: Recruitment, anonymization, and data-ownership practices that let lay people contribute ethically; “Relief is a systems variable; comfort and dignity override data collection.”
- Fail patterns worth looking for: Expanding variance with age, ledgers that drift steadily negative or positive without compensation, or shock responses that show no counterweight—patterns that would challenge LoF.

Subsections:

- 13.1 Longitudinal HCI, Practicalities
- 13.2 Compression Near the End
- 13.3 Life Events as Ledger Shocks
- 13.4 Research Notes: Missingness and Hierarchical Models
- 13.5 Citizen Science Without the Creepiness
- 13.6 Fail Patterns: Expanding Variance

Part VII — Rival Explanations, Fairly Presented

Chapter 14 — The Hedonic Treadmill and Opponent Processes

Promise: Presents hedonic adaptation and opponent-process theory at their strongest, then states precisely what they can and cannot guarantee compared with LoF. Lays out clean tests that distinguish statistical tendencies from a lifetime balance law.

Outcomes:

- Best evidence acknowledged fairly: Classic rebounds after windfalls and injuries, short- and mid-term moderation, and known biological bases of opponent processes are summarized without dismissal.
- Where rivals succeed vs. where they stop: These frameworks explain typical rebounds but do not guarantee a neutral lifetime ledger for each unified stream, nor do they specify end-of-life compression or horizon effects.
- Specific contrasts to test: LoF would predict horizon-dependent intensification and terminal variance drop; adaptation alone is silent on those signatures and allows permanent imbalance if environments are extreme.
- Tendency vs. Law (Research Note): We formalize why population-mean return is not equivalent to a per-stream guarantee and outline preregistered comparisons (naming one OOS metric once per chapter) to prevent moving goalposts.
- If Rivals Win: If all balancing phenomena reduce to adaptation and opponent dynamics with no residual LoF signatures, LoF collapses into these rivals; otherwise, rivals become mechanisms operating under a stricter balance constraint.

Subsections:

- 14.1 Best Evidence For
- 14.2 Where They Shine
- 14.3 What They Cannot Guarantee
- 14.4 Research Notes: Tendency vs. Law
- 14.5 If Rivals Win, What LoF Learns

Chapter 15 — Predictive Coding and Free-Energy

Promise: Explains predictive coding (PC) cleanly, then separates “minimize prediction error” from LoF’s lifetime balance requirement. Clarifies how to tell PC-only accounts from a true balance constraint.

Outcomes:

- The big idea, without teleology: A predictive system minimizes prediction error (free energy) by updating internal models or selecting actions that reduce sensory-model mismatch—no “wants/hates” phrasing.
- Affect as prediction error: Negative feelings map to high mismatch; relief maps to successful resolution—stated precisely and without implying a fairness guarantee.
- Why viability ≠ fairness: A system can minimize surprise yet sustain misery if misery is predictable; LoF concerns the integral of felt experience, not merely error reduction.
- Overlap vs. independence tests: We propose preregistered signatures (e.g., QS-residuals, end-of-life compression) and evaluate whether they are independent of canonical prediction-error circuits.
- If Rivals Win: If every balancing effect is fully reinterpretable as uncertainty reduction with no residual LoF signatures, LoF becomes a special case of PC; if strict terminal neutrality appears beyond PC, a balance constraint must be added to predictive models.

Subsections:

- 15.1 The Big Idea (Uncertainty Minimization)
- 15.2 Affect as Prediction Error
- 15.3 Why Viability ≠ Fairness
- 15.4 Research Notes: Overlap vs. Independence Tests
- 15.5 If Rivals Win, What LoF Learns

Chapter 16—Reinforcement Learning and Homeostasis

Promise: Examines RL and homeostatic regulation as explanations for behavior and affect, and tests whether optimization or equilibrium maintenance can deliver a per-stream balance guarantee. Considers hybrid models honestly.

Outcomes:

- Rewards, set points, and care: RL maximizes expected reward; homeostasis maintains set points; both can create stability without closing a lifetime hedonic ledger.
- Optimization isn't balance: An agent could achieve high cumulative reward while avoiding punishment, yet LoF forbids unbalanced ledgers at death of mind; the difference is empirical and testable.
- Composite rivals (hybrids): Predictive coding, RL, opponent processes, and regulation might approximate balance together; LoF treats them as mechanisms that may operate under stricter constraints.
- Model comparison and adversarial fits (Research Note): We outline preregistered fits and a single named OOS metric to compare RL/homeostasis against LoF-signature models, including collaboration with friendly skeptics.
- If Rivals Win: If pure RL/homeostasis explains data with no residual LoF signatures and no terminal compression, LoF is unnecessary; if LoF signatures persist, a balance constraint adds explanatory power.

Subsections:

- 16.1 Rewards, Set Points, and Care
- 16.2 Optimization Isn't Balance
- 16.3 Composite Rivals (Hybrids)
- 16.4 Research Notes: Model Comparison and Adversarial Fits
- 16.5 If Rivals Win, What LoF Learns

Part VIII — Evolution and Simulated Worlds

Chapter 17 — Natural Selection Meets a Law

Promise: Investigates how a balance constraint would coexist with evolution. Derives predictions across species and cultures without teleology and specifies evolutionary-grade tests that could falsify LoF.

Outcomes:

- Constraints the genome can't break: If LoF holds, selection can't produce organisms whose unified streams run permanently net-positive or net-negative in felt experience; we explain what this means in practice.
- Why control systems resemble QS: Evolved regulators (hunger, pain, pleasure) look like guardrails that prevent runaway states; we identify where these resemble a Queue System rather than a purpose-driven design.
- Cross-species predictions: Social mammals may show end-of-life calming and horizon effects; very simple agents may not—yielding comparative tests for LoF-consistent regulation versus mere fitness maintenance.
- Fitness-neutral vs. constraint-binding experiments (Research Note): We propose tests where survival fitness is held constant while emotional balance is manipulated, to avoid conflating fairness with viability.
- Fail patterns with systematic imbalance: If a lineage, species, or engineered strain reliably accumulates net suffering or net pleasure without counterweights and without survival penalties enforcing balance, LoF would be in trouble.

Subsections:

- 17.1 Constraints the Genome Can't Break
- 17.2 Why Control Systems Resemble QS
- 17.3 Cultural Echoes: Karma, Justice, Penance
- 17.4 Cross-Species Predictions
- 17.5 Research Notes: Fitness-Neutral, Constraint-Binding
- 17.6 Fail Patterns: Species with Systematic Imbalance

Chapter 18 — If Life Is a Game

Promise: Treats the “simulation/game” idea as an engineering thought experiment: if one were building a world with agents and emotions, would a balance constraint make the system stable and fair at the experiential level? We separate design logic from metaphysics and show how to test the analogy in code and in complex systems.

Outcomes:

- Why a designer would choose a constraint: Constraints prevent runaway misery/pleasure more cleanly than ad-hoc fixes, reduce exploit loops, and create predictable guardrails for agents—without implying cosmic intent.
- Constraints beat patches: We compare “balance by rule” to endless manual tweaks (deus-ex-machina rewards, punitive nerfs) and argue that a constraint is the scalable, elegant solution if experiential stability is required.
- Dreams as offline balancing passes: Revisit the book’s dream hypothesis through the game lens—low-cost “night work” that rebalances ledgers between rounds—stated as a testable mechanism, not a story.
- Research Notes: no-neutrality-by-fiat in code: In simulations, neutrality is not hard-coded as an outcome; instead, we add or withhold a balance constraint and predict distinct data signatures (variance compression, horizon effects).
- Indirect evidence from complex worlds: We look for natural systems that lack similar constraints and diverge, versus systems that remain stable with constraint-like feedback—offering analog clues without overreach.
- Fail patterns in simulation studies: If agent-based worlds without a constraint remain stable and life-like, or if adding a constraint produces contradictions instead of LoF signatures, the design analogy weakens.

Subsections:

- 18.1 Why a Designer Would Bake in LoF
- 18.2 Constraints Beat Patches (Cost and Elegance)
- 18.3 Dreams as Offline Balancing Passes
- 18.4 Research Notes: No-Neutrality-by-Fiat in Code
- 18.5 Indirect Evidence: Worlds That Fail Without Constraints
- 18.6 Fail Patterns in Simulation Studies

Part IX — Ethics and Human Dignity

Chapter 19 — What This Never Justifies

Promise: Sets non-negotiable ethics: LoF, if true, never licenses neglect, harm, or delay of relief. We enumerate duties for caregivers, researchers, and communicators so compassion and dignity always come first.

Outcomes:

- No license to ignore pain: Belief in LoF cannot excuse inaction; care must address suffering now, regardless of any longer-term balancing hypothesis.
- Duties of caregivers and researchers: Provide timely comfort, obtain informed consent, respect privacy, and stop or redesign any procedure that adds distress without benefit; observation follows care, not the reverse.
- Justice aimed at restoration: If balance mechanisms exist, justice should emphasize repair and rehabilitation over retribution, while keeping decisions evidence-guided and conservative at the margins.
- Non-sentience in simulation: Experiments about LoF must avoid sentient suffering; use non-sentient agents or retrospective/humane datasets to study mechanisms safely.
- Communication ethics: Avoid deterministic slogans (“it all balances out”), avoid euphemisms, present uncertainty, and never promise comfort that data do not support.
- Hard lines we will not cross: “Relief is a systems variable; comfort and dignity override data collection.” No deceptive withholding of care, no invasive protocols without consent, no coercion, and no “forced balance.”

Subsections:

- 19.1 No License to Ignore Pain
- 19.2 Duties of Caregivers and Researchers
- 19.3 Justice Aimed at Restoration
- 19.4 Research Notes: Non-Sentience in Simulation
- 19.5 Communication Ethics
- 19.6 Hard Lines We Will Not Cross

Chapter 20 — Hope, Freedom, and Daily Life

Promise: Explores how to live with the hypothesis—true or false—without fatalism. We show how meaning, choice, and responsibility remain intact inside constraint-style guardrails.

Outcomes:

- Freedom inside guardrails: Constraints limit outcomes, not agency; people still choose paths and knowing guardrails exist can reduce panic and support wiser action.
- Meaning without illusions: Value does not require cosmic purpose; purpose can arise from restoring balance in ourselves and others and from practical care that reduces open “tickets.”
- If the law is true: Expect cautious optimism and resilience; palliative care and social policies can align with natural balancing, not replace it.
- If the law is false: Our duty to relieve suffering increases precisely because there is no automatic counterweight—an actionable stance, not despair.
- Talking with skeptics: Frame LoF as a testable constraint, invite adversarial collaboration, and welcome disconfirming data; the tone is scientific humility, not dogma.
- A one-page summary to share: Provide a concise handout of the claim, the evidence plan, and the ethics so readers can communicate the idea clearly and responsibly.

Subsections:

- 20.1 Freedom Inside Guardrails
- 20.2 Meaning Without Illusions
- 20.3 If the Law Is True
- 20.4 If the Law Is False
- 20.5 Talking with Skeptics
- 20.6 A One-Page Summary to Share

Part X — Applying the Law: From Habits to Labs

Chapter 21 — The Ledger Gym: Habits, Queue Traps and Repair

Promise: A practical, evidence-minded handbook for living with the Law of Fairness day-to-day. You'll translate the core constructs (H , C , A , λ , Φ) into levers you can actually pull, build a simple "daily kit" that raises Φ and stabilizes λ , avoid the seven common queue traps (de-moralized, de-jargonized), and run a 14-day N-of-1 to see what truly helps you—ethically, measurably, and without teleology.

Outcomes:

- Understand why practice works: habits expand the admissible set A , raise flexibility Φ , and stabilize the shadow price λ ; uncontrolled pleasure/avoidance narrows A , lowers Φ , and destabilizes λ .
- Use the Daily Protocols with "what to do / what to measure / fail-pattern / high- Φ swap" for each.
- Spot the Seven Queue Traps (neutral names that map cleanly to folk "sins") and replace each with a practical, high- Φ alternative; includes a one-page table.
- Apply Social Guardrails: apology, reversibility, and reading "signals vs. scripts" so relationships function as healthy channels C rather than recurring traps.
- See Faith Practices as Queue Hygiene (e.g., fasting, sabbath, confession) with respectful, secular analogs—no teleology, just working guardrails.
- Run a 14-day N-of-1 using HCI-lite and simple stats (with a dispersion check) to learn what helps you; name success/fail criteria in advance.
- Know what not to do: over-control, moralizing, or using LoF to justify harm; keep the ethics line visible ("relief is a systems variable").

Subsections:

- 21.1 Warm-Up: Your Life as a Ledger Gym
- 21.2 Daily Habits as Guardrails and Repairs
- 21.3 Seven Queue Traps (Classical "Sins" Reimagined)
- 21.4 Social Relationships: External Guardrails and Repair Kits
- 21.5 Ritual and Reflection: Spiritual Practices as Queue Hygiene
- 21.6 N-of-1 Experiment: Training Your Own Guardrails
- 21.7 What Not to Do: Ethical and Practical Warnings

Chapter 22 — The Scientific Playbook

Promise: A comprehensive blueprint for full-scale Law-of-Fairness tests: audit-proof preregistration, reproducible HCI code and open data hygiene, multi-site replication with invariance checks, red-team challenges and bounties, and equivalence testing for terminal neutrality $L(T)$ —plus appended classroom kits so small studies can roll up cleanly into professional programs.

Outcomes:

- Full prereg packages that bind: locked hypotheses tied to LoF constructs ($H, C, A(t; \hat{H}, H, C), \lambda, \Phi$), multiverse options declared up front, deviation logs, and transparent pass/fail gates (e.g., neutrality criteria).
- Reproducible pipelines: versioned code, containerized analyses, hashed IDs / k-anonymity, and shareable de-identified or synthetic datasets so anyone can rerun end-to-end.
- Multi-site replication: harmonized protocols, measurement-invariance checks, federated or pooled analyses without scale drift; cross-cultural extensions.
- Adversarial review: red-team challenges, pre-registered bounties, and rival-model contests (e.g., lower predictive log-loss/WAIC) to surface hidden failure modes.
- Equivalence testing for $L(T)$: TOST/ROPE-style procedures with autocorrelation handled to confirm end-of-life neutrality bounds (mean $\pm 0.15 z$; slope $\pm 0.05 z/day$; variance ratio ≤ 0.80).
- Appended classroom kits: dream counterweights, a simple horizon task, teen-appropriate ethics and blinds, and consent templates that scale from class to clinic.

Subsections:

- 22.1 Full Prereg Packages
- 22.2 HCI Code and Open Data Hygiene
- 22.3 Multi-Site Replication
- 22.4 Red-Team Challenges and Bounties
- 22.5 Research Notes: Equivalence Testing for $L(T)$
- 22.6 Classroom Dream Counterweights
- 22.7 A Simple Horizon Task
- 22.8 Ethics and Blinds for Teens
- 22.9 Research Notes: Consent Templates

Part XI — The Case in One Place

Chapter 23 — The Ten Hardest Objections (and Our Answers)

Promise: Presents the strongest critiques of LoF and answers them without strawmen or rhetoric. The point is to map real failure conditions so the law can be confirmed or rejected in good faith.

Outcomes:

- Objections, head-on: From “You can’t measure feelings” to “It’s unfalsifiable,” each critique is stated in its sharpest form and answered with the book’s best evidence and tests.
- Distinguishing rivals: We explain why adaptation, predictive coding, and RL/homeostasis can be true and still fail to guarantee a neutral lifetime ledger, and what data would decide the matter.
- Ethical reassurance: We show why the idea never justifies neglect and how the ethics chapter constrains practice and public talk.
- Serious stress-test, not apologetics: Some objections may stand; if so, they limit or refute LoF—no moving goalposts.

Subsections:

- 23.1 “You Can’t Measure Feelings”
- 23.2 “Adaptation Explains It”
- 23.3 “You’re Moralizing Physics”
- 23.4 “Identity Is Fuzzy”
- 23.5 “Dreams Are Noise”
- 23.6 “The Brain Is a Prediction Machine”
- 23.7 “Simulations Prove Nothing”
- 23.8 “Evolution Wouldn’t Select This”
- 23.9 “It’s Unfalsifiable”
- 23.10 “It’s Dangerous to Say Suffering Balances”
- 23.11 Research Notes: Where to Find the Evidence

Chapter 24 — If Fairness Is Real

Promise: Closes with the discovery claim and a practical call to test, not believe. We distill what changes—for science, care, and daily life—if a balance constraint on felt experience really governs each unified stream.

Outcomes:

- The discovery claim: Under LoF the system must close each lifetime ledger at the death of mind; we restate the evidence path and constraints that make this claim empirical.
- What ordinary people can do: Keep channels open for relief, seek help, offer kindness, and support community practices that lower the real-world cost of counterweights; small acts matter regardless of the law's truth.
- A call for courage: Pre-register your own hypotheses, collaborate with rivals, and publish what you find; the invitation is to do the work, not salute a slogan.
- A structural design test: contrasting External Forced Balance with Internal Moderation-First societies to show how fairness-aware systems stabilize suffering without drifting into coercion, collapse, or hidden ledger transfers.
- Final reflection: Whether LoF stands or falls, telling the truth and relieving suffering are never wasted efforts; we end with a bridge to future tests.

Subsections:

- 24.1 The Discovery Claim
- 24.2 What Ordinary People Can Do
- 24.3 A Call for Courage (to Test, Not Believe)
- 24.4 Utopia Thought Experiments: External vs. Internal Moderation
- 24.5 Bridge to Synthesis

Chapter 25 — Final Synthesis

Promise: Bring every thread together into one definitive view of the Law of Fairness. Chapter 25 is the culmination of the entire work — a synthesis of science, philosophy, and ethics that tests whether life's symmetry of joy and sorrow can stand as a genuine law. Here, all disciplines converge — physics, psychology, systems theory, spirituality, and social ethics — to evaluate LoF's reach, its rivals, and its meaning. The goal is to leave the reader with a unified framework: what the law claims, how it might operate, what remains uncertain, and why the pursuit itself matters regardless of the outcome.

Outcomes:

- Re-states the Law of Fairness in its final form, linking its six assertions, core equations, and boundary conditions into one coherent model of lifetime equilibrium.
- Bridges metaphysical, physical, psychological, spiritual, and societal perspectives — showing how each lens contributes evidence or counter-argument to the central claim.
- Connects the Queue System, compensability $\Phi(\hat{L}, H)$, horizon λ_t , and the Hedonic Composite Index (HCI) into a single explanatory chain from experience to data.
- Clarifies how LoF coexists with moral agency — affirming compassion, responsibility, and the principle that relief and dignity override dogma or determinism.
- Explores how a fairness-aware worldview could reshape justice, care, and policy — inspiring cultures that actively promote balance rather than passively expecting it.
- Concludes that whether LoF proves true or false, the inquiry itself unites scientific rigor with moral purpose — inviting future generations to keep testing, refining, and practicing fairness through knowledge and kindness.

Subsections:

- 25.1 The Ontological and Metaphysical Perspective
- 25.2 The Physical and Systems Perspective
- 25.3 The Psychological Perspective
- 25.4 Spiritual and Moral Parallels
- 25.5 Societal and Ethical Implications
- 25.6 Final Reflections

Glossary of Key Terms

ACC (Anterior Cingulate Cortex): A cortical region associated with conflict monitoring and cost/effort evaluation. In LoF tests it is treated as a candidate hub where horizon-weighted compensability signals (Φ) may be detectable, without implying ACC “enforces” fairness.

Admissible Action Set ($\mathcal{A}(t)$): The subset of actions available at time t that keep eventual ledger neutrality feasible. Actions that would almost certainly make a neutral final ledger unreachable are excluded.

Admissible Policy: A decision rule (policy) whose trajectories remain within the admissible set over time, preserving terminal neutrality under the Queue System model.

Adversarial “Gaming”: Attempts—intentional or incidental—to exploit measurement rules so that HCI or $\bar{L}(t)$ shifts without corresponding changes in felt experience. Treated as a fail condition that pauses inference and triggers redesign.

Adversarial Fits (Rival Model Challenge): A preregistered comparison in which rival models are optimized on the same dataset to determine whether LoF-specific signatures persist after strong alternatives are fit.

Affective “Tickets”: Unresolved affective loads carried forward from salient experiences (e.g., a stressful day). Operationalized only through measurable signals; not inferred from narrative interpretation alone.

Affective Neuroscience (as used here): A mechanistic framing layer for HCI channels (brain, body, behavior), used descriptively to generate measurable predictions rather than as proof of LoF.

Agency: The capacity to choose among available options. LoF constrains feasible trajectories but does not override agency or imply moral desert.

AI (Artificial Intelligence): Non-biological systems that may qualify as unified streams only if preregistered integration and conscious-access thresholds are met.

Analgesia: Pain-relieving intervention. Relief is never withheld; analytically modeled as an exogenous window rather than manipulated for ledger testing.

Anesthesia: Medically induced loss of consciousness. Treated as a ledger pause when unity falls below threshold.

Anonymization: Privacy-preserving procedures including de-identification, hashed IDs, and k-anonymity.

Arousal: The intensity/activation dimension of emotion (calm → excited). Orthogonal to valence in core affect models.

Audit (Third-Party Audit): Independent review of preregistration compliance and analytic transparency.

Autocorrelation: Time-dependence in repeated measures; must be modeled when testing equivalence at closure.

Autonomic Physiology (Channel): Physiological measures (e.g., HRV, pupil dilation) used in HCl and horizon tasks.

Baseline (Matched Baseline): A preregistered reference period used for computing compression ratios and terminal comparisons.

Bayesian Hierarchical Model: Multi-level modeling framework used in longitudinal telemetry and missingness handling.

Best-System Law: A philosophical view of laws as summaries of regularities rather than governing teleological forces.

Biological Death: Organism death. LoF evaluates closure at “death of mind,” not merely biological cessation.

Blinding (Blinds): Procedures preventing expectancy bias (e.g., blinded coding, blinded adjudication).

Brain Signals (Channel): Neural measures (EEG, fMRI) used as HCl inputs and candidate QS correlates.

Calibration: Mapping raw channel measures to a common latent HCl scale with uncertainty.

Calibration Ladder: Within-person → between-person → cross-cultural calibration sequence.

Canonical Statement (LoF): For unified stream $0 \rightarrow T$, $L(T) = \int_0^T F(t) dt$ falls within preregistered neutrality bounds at closure.

Channels (Measurement Channels): Self-report, physiology, brain signals, behavior, dreams.

Cherry-Picking: Selective reporting avoided through preregistration and audits.

Choice Set: Full set of available actions at time t ; distinct from admissible set $A(t)$.

Chills Anchor: Shared affective anchor (e.g., music-induced chills) for calibration where ethically safe.

Closure (Ledger Closure): Condition that final ledger falls within neutral tolerance.

Cohort Endpoint (Shared Horizon): Shared deadlines producing population-level shadow price effects.

Compensability (Φ): Predicted change in probability of neutral final ledger if action u is taken.

Compensation (Counterweight): Measurable shift reducing imbalance.

Compression (Terminal Compression): Preregistered narrowing of mean/variance near end-of-life.

Configural Invariance: Same factor structure across groups.

Conflict (C): Competing alternative signal modeled as nuisance predictor in QS tests.

Confirmatory Factor Analysis (CFA): Latent-variable method validating HCl structure.

Consent (Informed Consent): Participant authorization; care precedes data.

Control Hubs: Candidate neural regions implicated in QS-like pruning (rlFG, ACC, vmPFC, insula).

Convex Terminal Loss ($\phi(L_T)$): Penalty increasing with imbalance; motivates rising λ_t in horizon formalism.

Core Affect: Two-dimensional valence/arousal framework underlying HCl.

Costate (λ_t): Shadow price multiplier on terminal imbalance in control-theoretic framing.

Counterfactual: Preregistered comparison distinguishing compensation from regression.

Cross-Validation: Held-out predictive testing to prevent overfitting.

Cultural Scripts vs Signals: Distinction between narrative expectations and measurable trajectories.

dACC: Dorsal ACC; urgency/conflict region candidate.

Death of Mind (T): Irreversible cessation of unified conscious access.

Decision Weight ($\omega(u; t)$): Softmax probability weight proportional to $\exp[\beta \cdot \Phi]$.

Delta (Δ): Change operator; first differences used in $\text{HCl}\Delta$.

Differential Ledger: $dL/dt = F(t)$.

Dream Ledger (Dream Counterweights): Hypothesized REM-based low-cost compensatory mechanism; testable and falsifiable.

Drift (Non-Stationary Drift): Measurement instability invalidating inference.

EEG: Electroencephalography; HCl channel and sleep-stage tool.

Endgame: Period when H_t is short and λ_t increases.

Equivalence Bounds: Preregistered tolerance band around zero.

Equivalence Testing (TOST): Two One-Sided Tests confirming neutrality within bounds.

Error Budget: Separation of latent ledger $L(t)$ and measured $\hat{L}(t)$ with uncertainty accounting.

Experiential Fairness: Fairness defined purely in felt experience.

Fail Pattern: Preregistered empirical signature that would count against LoF.

Federated Analysis: Multi-site analysis without centralizing raw data.

Felt Experience: Subjective pleasure/pain operationalized via HCl.

First Differences (Δz_k): Standardized change in channel k at time i .

Free Energy: Prediction-error minimization concept in rival PC models.

HCI (Hedonic Composite Index): Latent momentary net affect measure combining multiple channels.

$\text{HCl}\Delta$: First-difference formulation of HCl.

HCU (Hedonic Composite Unit): Unit of integrated HCl over time.

Hedonic Adaptation: Tendency to return toward baseline; rival explanation.

Horizon (H_t): Expected remaining lifetime $E[T - t]$.

\hat{H} (Estimated Horizon): Subjective/model-based estimate of horizon.

HRV: Heart-rate variability; effort/urgency marker.

Identity Rules (Pause/Split/Merge): Preregistered adjudication rules for unified streams.

Integrated Information and Control: Observable integration markers defining unity.

Invariance Ladder: Configural → Metric → Scalar testing sequence.

IRT (Item Response Theory): Latent modeling method for questionnaire scaling.

k-Anonymity: Privacy protection threshold.

Last-Window Behavior: Targeted compensatory action near endgame.

Latency (Pause): Period without ledger accrual due to absent consciousness.

Latent Ledger ($L(t)$): True cumulative $\int F(t) dt$.

Ledger Shock: Large positive/negative perturbation tested for compensation.

Lifetime Ledger (L): $\int_0^T F(t) dt$.

Log Loss: Predictive scoring rule for rival model comparison.

Longitudinal Telemetry: Repeated HCI measurement over months/years.

Measurement Invariance: Cross-group comparability condition.

Mean/Slope/Variance Gates: Preregistered end-of-life compression metrics.

Mixed-Effects Model: Model handling repeated measures and heterogeneity.

N-of-1: Single-subject experimental protocol.

NREM: Non-REM sleep; treated as negligible ledger accrual relative to REM/wake.

OFC: Orbitofrontal cortex; valuation region candidate.

Optional Stopping: Flexible stopping avoided via preregistration.

Out-of-Sample Metric (OOS): Predictive performance metric (e.g., WAIC, log loss).

Pause (Ledger Pause): Temporary halt of ledger accrual.

Physiology (Channel): Bodily measures in HCI.

Population Shadow Price: Cohort-level λ_t effect under shared horizons.

Power Analysis: Sample size justification.

Predictive Coding (PC): Rival framework minimizing prediction error.

Preregistration: Locking hypotheses/analysis before data.

Proxy: Measurable stand-in for latent variable.

QS (Queue System): Hypothesized mechanism pruning choices toward compensability.

QS-Residuals: Remaining variance after nuisance controls.

REM (Rapid-Eye-Movement Sleep): Dream-rich sleep stage evaluated as candidate counterweight context.

Replication (Multi-Site): Cross-site harmonized testing.

rIFG: Right inferior frontal gyrus; inhibitory control region candidate.

ROPE: Region of Practical Equivalence.

Scalar Invariance: Equality of intercepts across groups.

Shadow Price (λ_t / β): Horizon-dependent multiplier on compensability.

Softmax: Choice rule converting scores into probabilities.

State-Space Model: Time-series model estimating latent HCl with uncertainty.

State Vector ($S(t)$): Drive-state vector in $F(t) = -W \cdot dS/dt$.

Stopping Time (T): Operational closure time.

TOST: Two One-Sided Tests for equivalence.

Unified Conscious Stream: Continuous conscious flow treated as unit of analysis.

Unity Index (θ): Threshold-based measure of conscious unity.

Uncertainty Bands: Interval estimates around HCl and $\bar{L}(t)$.

Valence: Positive/negative quality of experience.

Valence Inversion: Dream prediction of opposite-valence compensation.

Variance Ratio: Terminal variance relative to baseline (e.g., ≤ 0.80).

vmPFC: Ventromedial prefrontal cortex; valuation region candidate.

W (Drive-Weight Vector): Weight vector in $F(t) = -W \cdot dS/dt$.

WAIC: Widely Applicable Information Criterion for model comparison.

β (Shadow-Price Parameter): Softmax multiplier rising as horizon shrinks.

ϵ (Neutral Tolerance Notation): Symbolic tolerance around zero; operationalized empirically via half-width K in HCU.

Notation and Mathematical Conventions

This book uses a mix of continuous-time math (integrals, stochastic processes) and discrete-time operationalizations (sampled measurements, day-level summaries). Unless explicitly stated otherwise, the conventions below apply throughout.

Mathematical typography and indexing conventions

- **Time**

- **t** — Continuous time variable (e.g., seconds, minutes, hours).
- **t_i** — Discrete sampling time index ($i = 1, 2, \dots$), used when data are collected at irregular or regular intervals.
- **Δt_i** — Time step between samples ($t_i - t_{i-1}$).
- **T** — Terminal time (death of the mind / end of unified conscious access), i.e., the end boundary of the ledger integral.

- **Indices**

- **i** — Individual (person) index in group/multi-person data contexts.
- **k** — Channel / indicator index (e.g., HRV, self-report item, EEG feature).
- **r** — Resource type index (Queue System resources: sleep, time, money, social support, etc.).
- **d** — Day index in day-level summaries (e.g., ΔL_d).
- **g** — Group index (e.g., culture/site/subpopulation) in measurement invariance and multigroup models.
- **s** — Posterior draw / Monte Carlo sample index ($s = 1, \dots, S$).
- **b** — Branch index in “branching stream” thought experiments ($b = 1, \dots, B$).

- **Hats, bars, and window operators**

- **$\hat{}$ (hat)** — Estimated / inferred / model-based quantity (e.g., $\hat{L}(T)$, \hat{H}_t , $\hat{HCI_forecast}$).
- **\check{X}** (X with “window mark”) — A window-averaged or window-summary statistic used operationally (e.g., \check{HCI}_{last} as “average HCI in the final window”).

- ΔX — Discrete difference or change (context-dependent; see Δ entries below).
- x^- (superscript minus in the text, as in $\Delta^- L$) — “Deficit-side” / negative-side quantity as used in the book’s QS and dream-balancing operationalizations (e.g., “recent negative imbalance”); when formalized, it is the “negative-part” style usage (deficit side) rather than the positive side.
- **Core operators and relations**
 - \int — Integral over time (continuous-time accumulation).
 - Σ / \sum — Summation.
 - d/dt — Time derivative.
 - $dX(t)$ — Differential increment; in SDEs, read in the Itô sense when paired with $dW(t)$.
 - $E[\cdot]$ — Expectation.
 - $Var(\cdot), Cov(\cdot)$ — Variance and covariance.
 - $Pr\{\cdot\}$ — Probability.
 - $|\cdot|$ — Absolute value (scalar magnitude); also used in bounds like $|L(T)| \leq K$.
 - \mathbb{R}, \mathbb{R}^d — Real numbers; d-dimensional real vector space.
 - \cdot — Dot product when used as $W \cdot dS/dt$.

- **Randomness and stochastic process notation**

- $(\Omega, \mathcal{F}, \mathbb{P})$ — Sample space, sigma-algebra, and probability measure.
- \mathcal{F}_t — Filtration (information available up to time t).
- $W(t)$ — Standard Brownian motion; $dW(t)$ is its increment.
- σ — Diffusion scale (noise magnitude) in SDEs.

Core Law of Fairness ledger and affect formalism

- **LoF (Law of Fairness)** — The central hypothesis: for a *unified conscious stream*, the terminal (end-of-mind) ledger closes to zero at the terminal boundary. Operational tests use preregistered equivalence bands and measurement gates rather than claiming perfect measurement of exact zero.

- **F(t)** — Instantaneous net affect (valence rate) at time t. In the state-change formalism, $F(t) = V(t) = -W \cdot dS(t)/dt$. Positive F(t) corresponds to net pleasure (drive reduction); negative corresponds to net pain (drive increase).
- **V(t)** — Felt valence rate at time t; by definition $V(t) = F(t)$ in the formalism used here.
- **L(t)** — True cumulative life ledger up to time t: $L(t) = \int_0^t F(\tau) d\tau$.
- **L(T)** — Final life ledger at terminal time T: $L(T) = \int_0^T F(t) dt$.
- **$\hat{L}(t)$** — Measured/estimated ledger up to time t (data-driven proxy): typically constructed by integrating a momentary measurement like HCl, i.e., $\hat{L}(t) = \int_0^t HCl(\tau) d\tau$ (continuous-time idealization) or its discrete approximation.
- **$\hat{L}(T)$** — Estimated terminal ledger derived from data and the measurement model; distinct from the latent “true” **L(T)**.
- **K** — Neutrality bound (half-width). A preregistered threshold defining the neutral band $[-K, +K]$ in HCU. A “neutral final ledger” operationally means $|\hat{L}(T)| \leq K$ (or the corresponding posterior probability of this event exceeds the preregistered gate).
- **HCU (Hedonic Composite Unit)** — Ledger unit (the integral of the momentary HCl measure over time). In practice, HCU is the unit used for **L(t)**, **$\hat{L}(t)$** , and the neutrality bound **K**.
- **S(t)** — Latent drive-state vector in \mathbb{R}^d (homeostatic and motivational drives) whose change generates affect in the state-change formalism.
- **d** — Dimensionality of the drive-state vector $S(t) \in \mathbb{R}^d$.
- **W** — Drive-weight vector (a fixed row vector of positive weights after calibration), defining how drive changes contribute to net affect via $-W \cdot dS/dt$.

Important overload note: This **W** is *not* Brownian motion **W(t)** and is *not* queue waiting time **Wq**.

HCI measurement model notation (channels → composite)

- **HCI (Hedonic Composite Index)** — A calibrated latent metric of momentary felt experience (with uncertainty), constructed from multiple channels (self-report + physiology + behavior where available). HCl is the per-time affect estimate that is integrated into HCU.

- **HCI Δ (delta-based HCI)** — Discrete-time operational momentary affect: $\text{HCI}(t_i) = \sum_k w_k \Delta z_k(t_i)$, where Δz_k is the first difference of the standardized channel k between t_{i-1} and t_i . (In the text this is sometimes written as $\text{HCI}(t_i)$ when the Δ construction is implicit.)
- **$z_k(t)$** — Standardized (z-scored / normalized) feature for channel k at time t .
- **$\Delta z_k(t_i)$** — First difference of channel k at sample i (discrete-time change).
- **w_k** — Channel weight for channel k in the composite index.
- **$y_k(t)$** — Observed measurement for channel k at time t (generic observation notation).
- **$h_k(\cdot)$** — Observation/measurement function mapping latent affect to observed channel space in state-space formulations.
- **$\varepsilon_k(t)$** — Measurement noise for channel k (and/or residual term in the measurement equation).
- **$\check{\text{HCI}}_{\text{last}}$** — Window-summary HCI statistic in the final preregistered window near T (e.g., a mean/aggregate of HCI over the “last window” used for end-of-life closure diagnostics).
- **$\hat{\text{HCI}}_{\text{forecast}}$** — Model-based forecast/prediction of the final-window HCI statistic (used when comparing predicted vs observed end-window behavior).

Queue System decision model notation ($\Phi, \omega, \text{horizon, resources}$)

- **QS (Queue System)** — Proposed mechanism that biases decisions toward actions that improve feasibility of terminal ledger neutrality as the horizon shortens. It is modeled as a decision bias, not as an override of agency.
- **u** — An action (choice option).
- **u_t** — Action taken at time t .
- **π** — A policy (a rule mapping history/state to actions over time).
- **π_u** — A policy variant that “forces” action u at time t (used in counterfactual definitions of Φ).
- **$\mathcal{U}(t)$** — Universe of logically possible actions at time t .
- **$\mathcal{A}(t)$** — Admissible actions at time t (feasible given constraints).

- $\mathcal{S}(t)$ — Selectable actions at time t (the actual menu of actions the agent can choose, typically a subset of admissible actions given attention, cognition, context, etc.).
- H_t — Horizon proxy at time t (remaining time / subjective remaining life / remaining “distance” to terminal boundary). Shorter H_t corresponds to being “closer” to T .
- $\Phi(u; L(t), H_t)$ — Feasibility of compensation for action u at time t . In the book’s operational definition, Φ is the expected *increment* in the probability of terminal neutrality if u is taken (relative to a baseline policy), optionally net of horizon-weighted resource shadow prices:

$$\Phi(u; L(t), H_t) = \Delta \Pr(|L(T)| \leq K | \text{data}) - \sum_r \lambda_r(t) \Delta_r(u_t)$$

(with the understanding that the precise conditioning set and baseline are defined in the local experiment design).

- $\omega(u; t)$ — Decision weight / choice probability for action u at time t , defined via a softmax over Φ : $\omega(u; t) \propto \exp[\beta(H_t) \cdot \Phi(u; L(t), H_t)]$ with normalization across $u \in \mathcal{S}(t)$.
- $\beta(H)$ — Horizon-dependent “gain” / inverse-temperature controlling how strongly Φ affects choice as the horizon changes. In the book: $\beta(H) = \beta_0 + \beta_1 / (H + \delta)$, where $\delta > 0$ is a small constant preventing blow-up at $H = 0$.
- β_0, β_1 — Parameters in the horizon-gain function $\beta(H)$.
- ϵ (probability-gate tolerance) — Small tolerance used in preregistered probabilistic gates such as: $\Pr(|L(T)| \leq K | \text{data}) \geq 1 - \epsilon$.

Important overload note: ϵ is also used elsewhere as an error/innovation term (see below); subscripts/context disambiguate.

Resource, queueing, and congestion notation (QS “resources”)

- $R(t)$ — Resource budget vector at time t (e.g., sleep/time/money/social support), as used in QS resource-constraint discussions.
- $\Delta_r(u_t)$ — Resource decrement (or consumption) of resource r caused by taking action u at time t .
- $\lambda_r(t)$ — Shadow price (marginal value / marginal cost) of resource r at time t in Φ -style feasibility scoring. Larger $\lambda_r(t)$ means that spending resource r is more “expensive” in the QS objective.

- Λ_t — Population-level / global shadow price construct used in the text when discussing shared constraints across individuals (appears in the “population shadow price” framing).
- $\text{Cong}_r(t)$ — Congestion index for resource r at time t (a scalar proxy derived from queueing constructs such as expected waiting time).
- W_q — Expected waiting time in queueing notation (used as a proxy for congestion).
- λ_{arr} — Arrival rate in queueing examples (input rate into a queue).
- μ (queueing) — Service rate in queueing examples (output/service capacity).
- ρ (queueing) — Utilization ratio in queueing examples (commonly λ_{arr}/μ , with context-specific generalization).

Important overload note: ρ is also used as an AR coefficient in the dream model.

- c_r — Resource-specific scaling constant used to map queue quantities (e.g., waiting time) into a shadow-price-like scale (appears in $\lambda_r \approx c_r \cdot \hat{W}_{\{q,r,t\}}$ type approximations).
- $\hat{W}_{\{q,r,t\}}$ — Estimated queue wait / queue-derived proxy at time t for resource r (notation appears in the queueing-to-shadow-price mapping).

Horizon weighting and event-impulse notation

- $h(H)$, $g(H)$ — Generic monotone decreasing functions of horizon H used in horizon-weighting terms (e.g., scaling event impulses or neural signals by $1/H$ -like factors). The specific functional form is stated locally when used.
- t_j — Time of an event impulse j (event time points).
- η_j — Magnitude of an impulse at event j (e.g., a sharp affective event input in a simplified model).
- $\delta(t - t_j)$ — Dirac delta used to represent instantaneous impulses in continuous-time idealizations.

Important overload note: δ is also used elsewhere (e.g., for random-walk drift parameters); the Dirac delta is always written with an explicit argument ($t - t_j$).

Stochastic-process, martingale, and OST notation

- $\mu(t)$ — Predictable component (drift) in an SDE decomposition of instantaneous affect, as in $F(t) = \mu(t) + \varepsilon(t)$.
- $\varepsilon(t)$ — Mean-zero innovation term in continuous-time; in Itô formulations the innovation is modeled via $\sigma dW(t)$.

Important overload note: ε also denotes probability-gate tolerance in equivalence gates; subscripts/context disambiguate.

- σ — Diffusion scale in SDEs, controlling noise magnitude.
- $dW(t)$ — Brownian increment (innovation term) in continuous time.
- $\Psi(L, H)$ — Drift-regularization function used in the martingale framing (e.g., odd in L , increasing with $|L|$, growing as horizon shrinks), appearing in constructions like $\mu(t) = -\lambda \Psi(L(t), H_t)$ (local chapter notation).
- λ (SDE / martingale sections) — Nonnegative strength parameter scaling the restoring drift toward closure in martingale-style reasoning.

Important overload note: λ also appears as a dream-model coefficient and as a resource shadow price $\lambda_r(t)$; subscripts/context disambiguate.

- $M(t)$ — Transformed process constructed to have martingale properties under stated conditions (e.g., $M(t) = L(t) + \lambda \int_0^t \Psi(L(s), H(s)) ds$ in the chapter's construction).
- $(\Omega, \mathcal{F}, \mathbb{P})$ — Sample space, sigma-algebra, and probability measure in probability-theory statements.
- \mathcal{F}_t — Filtration (information available up to time t).
- τ — Stopping time (random time defined with respect to \mathcal{F}_t).
- **OST (Optional Stopping Theorem)** — The theorem relating $E[M(\tau)]$ to $E[M(0)]$ under standard conditions (e.g., bounded stopping time or suitable integrability conditions).
- **UI (uniform integrability)** — The uniform-integrability condition referenced as a sufficient condition for certain OST statements (appears explicitly as “UI” in the text).
- **SDE** — Stochastic differential equation.

- **OU (Ornstein–Uhlenbeck)** — Mean-reverting stochastic process family referenced in the book's stochastic modeling discussions.

Inference, testing, and equivalence-gate notation

- **H_0, H_1** — Null and alternative hypotheses.
 - H_0^-, H_0^+ — The two one-sided nulls used in TOST-style equivalence testing.
- **TOST** — Two One-Sided Tests procedure for equivalence testing.
- **ROPE** — Region Of Practical Equivalence (Bayesian analogue of an equivalence band).
- **SESOI** — Smallest Effect Size Of Interest (used to define equivalence/ROPE width operationally).
- **CI** — Confidence interval (frequentist).
- **Credible interval / HDI** — Bayesian posterior interval; **HDI** is the Highest Density Interval.
- **$p(\cdot)$** — Density or probability mass function; $p(\theta | y)$ denotes a posterior distribution, $p(y | \theta)$ a likelihood, and $p(\theta)$ a prior.
- **$Y_{0:t}$** — Observed dataset from time 0 through time T, used in expressions like $p(F(\cdot) | Y_{0:t})$.
- **α** — Significance level / Type I error rate (frequentist).
Important overload note: α also appears with subscripts in measurement equations as intercept-like parameters (α_k); subscripts/context disambiguate.
- **α -spending** — Sequential-testing control scheme referenced in the martingale/monitoring discussion (e.g., O'Brien–Fleming-style boundaries).
- **BF_{01}** — Bayes factor comparing H_0 to H_1 (appears in the text's Bayesian-testing discussion).
- **AIC, BIC, WAIC** — Model comparison/selection criteria referenced in the model evaluation table.
- **LOO** — Leave-one-out cross-validation (Bayesian/likelihood-based predictive evaluation context).
- **PSIS** — Pareto-smoothed importance sampling (used in PSIS-LOO).

- **MDL** — Minimum Description Length (information-theoretic model selection criterion referenced in the table).
- **PPC** — Posterior predictive check.
- **\hat{R} (R -hat)** — MCMC convergence diagnostic (Gelman–Rubin style), referenced in the inference checklist.
- **ESS** — Effective sample size (MCMC diagnostic).

Regression, neural signal, and QS-residual notation

- **A(t)** — Observed neural/physiological activity time series used in QS-residual regression examples (e.g., ROI signal).
- **y_ROI(t)** — Observed signal in a specified region of interest (ROI), used in neural-model equations.
- **Backbone(t)** — Baseline regressor (a nuisance/baseline component in neural models) used in examples.
- **RT(t)** — Reaction time (behavioral measure).
- **log(RT(t))** — Log-transformed reaction time used as a covariate.
- **U(t)** — Local utility term in the QS-residual regression example (not the same as the Unity Index term used elsewhere in prose).
- **C(t)** — Conflict/cost term in the QS-residual regression example.
- **Arousal(t)** — Arousal covariate term.
- **$\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$** — Regression coefficients in QS-residual regression (distinct from $\beta(H)$ inverse-temperature usage; context disambiguates).
- **$\epsilon(t)$** — Regression residual/noise term (distinct from equivalence-gate ϵ ; context disambiguates).
- **BOLD_ACC(t)** (and similar) — BOLD fMRI time series extracted from a given ROI (ACC shown as an example).

Dream / sleep balancing operational notation

- **REM** — Rapid-eye-movement sleep (dream-rich stage).
- **NREM** — Non-rapid-eye-movement sleep. In the book’s operational framing, *deep dreamless* NREM is treated as a period with Unity Index below threshold

where ledger accrual is paused (i.e., no HCU entries), while other sleep states are handled according to measured conscious access indicators.

- **D_d** — Dream affect index on day d (day-level dream affect variable in the AR-style model).
- **p** (dream model) — Autoregressive parameter in the dream model: $D_d = p D_{d-1} + \dots$ (distinct from ρ utilization in queueing).
- **λ** (dream model) — Dream-model coefficient coupling prior deficit to dream affect (distinct from λ_r and λ in SDE sections).
- **ε_d** — Day-level innovation/noise term in the dream model.
- **REM%** — Fraction of sleep time spent in REM (day-level covariate in the dream model).
- **ΔL_d** — Day-level ledger change, defined operationally in the Core Formulas section as: $\Delta L_d = \bar{L}(\text{morning}) - \bar{L}(\text{evening})$ (with sign conventions as stated locally).
- **$\Delta^- L_{d-1}$** — Prior-day deficit-side ledger quantity used as a predictor in dream-balancing models (as described in the text around the dream model).

Measurement invariance, CFA/IRT, and latent-variable notation

- **CFA** — Confirmatory factor analysis.
- **SEM** — Structural equation modeling.
- **IRT** — Item response theory.
- **$y_{\{g,t\}}^{(k)}$** — Observed indicator k at time t in group g (multigroup measurement model notation).
- **$F_{\{g,t\}}$** — Latent affect factor for group g at time t in CFA/IRT formulations.
- **$v_{\{k,g\}}$** — Item/indicator intercept for indicator k in group g.
- **$\lambda_{\{k,g\}}$** — Factor loading for indicator k in group g.

Important overload note: λ also appears elsewhere (shadow prices; drift strength; dream coefficient).

- **$\Gamma_{\{k,g\}}$** — Coefficient vector/matrix for covariates in the measurement model (e.g., nuisance regressors Z).

- $Z_{\{g,t\}}$ — Covariate vector in the measurement model (nuisance covariates).
- $\varepsilon_{\{g,t\}^{\{(k)\}}}$ — Residual/measurement noise for indicator k in group g at time t .
- Ψ_k — Residual variance term (indicator-specific), and Ψ_g — residual covariance matrix in multigroup SEM/CFA contexts.

Important overload note: Ψ also appears as $\Psi(L,H)$ drift-regularizer in the martingale framing.

- Λ_g — Loading matrix for group g in multigroup factor models.
- **DIF** — Differential item functioning (group-dependent item behavior).
- $a_{\{k,g\}}$ — IRT discrimination/slope parameter for item k in group g .
- $b_{\{k,gc\}}$ — IRT threshold parameter for item k , group g , category c .
- $\text{logit}(\cdot)$ — Logit link; $\text{logit}^{-1}(\cdot)$ is the logistic/sigmoid function (inverse-logit) used in graded response models.
- **CFI, RMSEA, TLI, SRMR** — Fit indices referenced in the measurement model/invariance discussion.
- **ΔCFI** — Change in CFI across invariance constraints (used as a practical invariance diagnostic in the text).

Consciousness-unity notation and disambiguations

- **Unity Index** — The book's operational construct intended to quantify whether conscious access is sufficiently unified to count as a single stream for ledger accounting (defined and discussed in Chapter 9.2). It is used to determine when ledger accrual is “on” versus “paused” (e.g., in deep dreamless NREM, certain anesthetic states).
- θ — Threshold parameter used with the Unity Index ($\text{Unity Index} \geq \theta$ indicates unified access; below θ indicates ledger pause under the operational rule).
- **IIT (Integrated Information Theory)** — A theory of consciousness referenced in the Unity Index discussion.
- **Φ_IIT** — IIT's integrated information measure, written with an IIT subscript in the book specifically to avoid confusion with LoF's Φ feasibility function.
- **GWT (Global Workspace Theory)** — Another consciousness theory referenced in the same context.

Neuroimaging / physiology / experimental-design acronyms used in LoF tests

- **ROI** — Region of Interest (predefined brain region used for analysis).
- **ACC** — Anterior cingulate cortex.
- **OFC** — Orbitofrontal cortex.
- **vmPFC** — Ventromedial prefrontal cortex.
- **rIFG** — Right inferior frontal gyrus.
- **BOLD** — Blood-oxygen-level-dependent fMRI signal.
- **fMRI** — Functional magnetic resonance imaging.
- **EEG** — Electroencephalography.
- **MEG** — Magnetoencephalography.
- **EOG** — Electrooculography.
- **EMG** — Electromyography.
- **ECG** — Electrocardiography.
- **PPG** — Photoplethysmography.
- **HRV** — Heart rate variability.
- **EDA / GSR** — Electrodermal activity / galvanic skin response.
- **TMS** — Transcranial magnetic stimulation.
- **tDCS** — Transcranial direct current stimulation.
- **N-of-1** — Single-subject experimental design (within-person intervention/measurement loop).

Data, modeling, and computation acronyms used in the book

- **AI** — Artificial intelligence.
- **LLM** — Large language model.
- **ML** — Machine learning.
- **RL** — Reinforcement learning.
- **AR(1)** — First-order autoregressive model (used explicitly in the dream model).

- **HMM** — Hidden Markov model.
- **GLM** — Generalized linear model.
- **GAM** — Generalized additive model.
- **PCA** — Principal component analysis.
- **ICA** — Independent component analysis.
- **MCMC** — Markov chain Monte Carlo.
- **HMC** — Hamiltonian Monte Carlo.
- **VI** — Variational inference.
- **MAP** — Maximum a posteriori estimate.
- **MLE** — Maximum likelihood estimate.

Symbol overload warnings

Some letters are reused in different chapters because they live in different “formal layers.” When they are reused, the book relies on subscripts and context:

- **W** (drive weights) vs **W(t)** (Brownian motion) vs **W_q** (queue waiting time).
- **λ** (drift strength in SDE) vs **λ_{r(t)}** (resource shadow price) vs **λ** (dream-model coefficient) vs **λ_{k,g}** (factor loading).
- **p** (AR coefficient in dream model) vs **p** (queue utilization).
- **ε** (equivalence-gate tolerance) vs **ε(t)** / **ε_{k(t)}** (innovation/error terms).

Core Formulas and Equations

Admissible Action Set (Queue System):

$$\mathcal{A}(t) = \{ u \in \mathcal{U}(t) : \Pr[L(T) \in [-K, K] | L(t), H_t, u] \geq 1 - \varepsilon \}.$$

The momentary action menu is restricted to choices that keep terminal neutrality sufficiently probable.

Baseline Latent Dynamics (null example):

$$F_{t+1} = \alpha F_t + \sum_k \theta_k X_{k,t} + \eta_t, \text{ with } |\alpha| < 1.$$

Balancing Drift (example):

$$\mu(t) = -\lambda \Psi(L(t), H_t).$$

Branching-Stream Ledger (thought experiments):

$$L(T) = \sum_{b=1}^B \int_{l_b} F_b(t) dt.$$

(Branch-weighting is specified locally when invoked.)

Channel Observation Model (generic):

$$y_{k,t} = h_k(S_t, (S_{t+1} - S_t)/\Delta t, F_t) + \varepsilon_{k,t}.$$

Choice Softmax (practitioner form; example):

$$\Pr[\text{choose } u] = \text{softmax}[\theta_U U(u) + \theta_C C(u) + \theta_\Phi \Phi(u) + \theta_H \Phi(u) H^{-1} - \sum_r \theta_{\lambda_r} \lambda_r \Delta_r(u)].$$

Compensability / Feasibility Score:

$$\Phi(u; L(t), H_t) = \Delta \Pr(|L(T)| \leq K | l_t, u) - \sum_r \lambda_r \Delta_r(u_t).$$

A compensability score combining neutrality-probability gain with resource shadow-price penalties.

Compensability (net form with explicit resource penalties; example):

$$\Phi_{\text{net}}(u; t) = \Phi(u; t) - \sum_r \lambda_r \Delta_r(u_t).$$

Compression Ratio (end-of-life variance comparison):

$$\rho \equiv \sigma_{\tau}(T) / \sigma_{\text{base}}.$$

(With $\kappa = 0.20$, the preregistered test is H0: $\sigma_{\tau,i}(T) \geq (1 - \kappa) \sigma_{\text{base},i}$ vs HA: $\sigma_{\tau,i}(T) < (1 - \kappa) \sigma_{\text{base},i}$)

Congestion Index (one example):

$$\text{Cong}_r(t) = z(\text{wait}) + z(\text{denials}) + z(\text{utilization}) - z(\text{idle_slack}).$$

Congestion-to-Price Mapping (examples):

$$\lambda_{r,t} = c_r \cdot \hat{W}_{q,r,t} \text{ or } \lambda_{r,t} = w_r \times \text{Cong}_r(t).$$

Cost-Penalized Decision Weight (example):

$$\omega(u; t) \propto \exp[\lambda_t \Phi(u; L_t, H_t) - \text{Penalty}(u, t)].$$

Cumulative Ledger:

$$L(t) = \int_0^t F(\tau) d\tau.$$

Differential Ledger:

$$dL/dt = F(t).$$

Doob Decomposition (drift + martingale; example framing):

$$L(t) = A(t) + M(t), \text{ with } A(t) = A_0(t) - \lambda \int_0^t \Psi(L(s), H_s) ds.$$

Dream Counterweight (illustrative):

$$E[D_{\text{night}} | \Delta L_{\text{day}}, H] = -\alpha(H) \Delta L_{\text{day}} + \eta.$$

Dream Compensation Model (day → night; illustrative):

$$E[D_{\text{night}} | \Delta L_{\text{day}}, H, R] = -\alpha(H) \cdot \Delta L_{\text{day}} + \eta.$$

Drive–Ledger Identity (state–change formalism):

$$\int_0^T F(t) dt = \int_0^T [-W \cdot dS(t)/dt] dt = -W \cdot (S(T) - S(0)),$$

when W is treated as constant over the interval.

Drift + Noise Decomposition:

$$F(t) = \mu(t) + \varepsilon(t).$$

Effective Sample Size under AR(1) (heuristic):

$$n_{\text{eff}} \approx n_{\text{pts}} \cdot (1 - \rho_{\text{AR}}) / (1 + \rho_{\text{AR}}).$$

Empirical Choice Model (example):

$$\Pr(u_t) = \text{softmax}(\theta_0 + \theta_U U(u_t) + \theta_{\text{Conf}} \text{Conf}(u_t) + \theta_{\Phi} \Phi(u_t) + \theta_H \Phi(u_t) H_t^{-1} - \sum_r \theta_r \lambda_r(t) \Delta_r(u_t)).$$

Equivalence (Neutrality) Test (TOST; frequentist framing):

$$H_0: L(T) \leq -K \text{ or } L(T) \geq +K; H_1: -K < L(T) < +K.$$

A Two One-Sided Tests (TOST) procedure is used on an estimated $\hat{L}(T)$ with its preregistered uncertainty model.

Example Endgame Gates (illustrative defaults used in preregistration templates):

Mean_ROPE_last ∈ [-0.15, +0.15] (z units),

Slope_ROPE_last ∈ [-0.05, +0.05] (z/day),

Var_end / Var_base ≤ 0.80.

Generic Channel Measurement Model:

$$y_k(t) = \alpha_k + \beta_k F(t) + \varepsilon_k(t), \text{ with } \varepsilon_k(t) \sim N(0, \sigma_k^2).$$

A generic measurement equation mapping latent affect to each observed channel.

Graded-Response IRT (ordinal outcomes):

$$\Pr(y_{k,t} \geq c | F_t) = \text{logit}^{-1}[a_k(F_t - b_{kc})].$$

HCI (delta-based operational definition):

$$\text{HCI}(t_i) = \sum_k w_k \cdot \Delta z_k(t_i).$$

At each discrete time, HCI is the weighted sum of standardized first differences from validated channels.

HCI Definition (level-based; alternate operationalization):

$$\text{HCI}(t_i) = \sum_k w_k \cdot z_k(t_i).$$

At each discrete time t_i , HCI is the weighted sum of standardized channels.

HCU (discrete-time total over an interval):

$$\text{HCU}[a,b] = \sum_{t_m \in [a,b]} \text{HCI}(t_m) \Delta t_m.$$

HCU Calibration Map (two-anchor linear scaling; illustrative):

$$\text{HCU}(t) = a F_t + b,$$

$$\text{with } a = (1 - (-1)) / (\Delta F_{\text{analgesia}} - \Delta F_{\text{cold}}),$$

$$\text{and } b = -a(\Delta F_{\text{analgesia}} + \Delta F_{\text{cold}})/2.$$

Horizon Estimation (proxy model; illustrative):

$$H_{t+1} = H_t - 1 + \varepsilon_t, \text{ and } y_k(t) = a_k H_t + \text{noise}.$$

Horizon Gain (example):

$$\beta(H) = \beta_0 + \beta_1 / H.$$

Horizon Gain Function (stabilized example):

$$\beta(H) = \beta_0 + \beta_1 / (H + \delta), \text{ with } \beta_1 > 0.$$

Horizon-Priority Adjustment (one implementation; example):

$$(\lambda_{r,t} - \kappa_r H_{t-1})^+.$$

Insula Imbalance Proxy (example):

$$\text{BOLD_insula}(t) = \beta_0 + \beta_1 H_{t-1} |L(t)| + \varepsilon.$$

Interrupted Time Series (pre/post; example):

$$\text{HCI}_t = \beta_0 + \beta_1 t + \beta_2 \cdot 1[t \geq t_0] + \beta_3 \cdot (t - t_0) \cdot 1[t \geq t_0] + \gamma^\top \text{Nuis}_t + \epsilon_t.$$

Ledger Costate (optimal control view; example):

$$\lambda_t = \partial J / \partial L.$$

Ledger SDE and Drift Cancellation (example):

$$dL(t) = \mu(t) dt + \sigma dW(t),$$

$$\text{with } \mu(t) = -\lambda \Psi(L(t), H_t, C_t).$$

Under this condition, the transformed process

$$M(t) = L(t) + \int_0^t \lambda \Psi(L(s), H_s, C_s) ds$$

$$\text{satisfies } dM(t) = \sigma dW(t).$$

Life-Ledger Integral (terminal):

$$L(T) = \int_0^T F(t) dt.$$

This definition of the lifetime ledger sums net affect (pleasure minus pain) over the entire conscious lifespan up to the operational stopping time T (“death of mind”).

LoF-Augmented Dynamics (example):

$$F_{t+1} = \alpha F_t + \sum_k \theta_k X_{k,t} + \lambda_t \Phi_t + \eta_t,$$

$$\text{with } \lambda_t = g(H_t) = H_t^{-1}.$$

Log Reaction-Time Model (example):

$$\log RT(t) = \beta_0 + \beta_1 \cdot \Phi(u_t; L(t), H_t) + \beta_2 \cdot \text{Conf}(t) + \varepsilon(t).$$

Martingale / SDE Framing (ledger; example):

$$dL(t) = \mu(t) dt + \sigma dW(t),$$

$$\text{with } \mu(t) = -\lambda \Psi(L(t), H_t).$$

Martingale Transform (under that drift; example):

$$M(t) = L(t) + \lambda \int_0^t \Psi(L(s), H_s) ds,$$

$$\text{so } dM(t) = \sigma dW(t).$$

Mean-Reverting Affect with Impulses (OU + events; illustrative):

$$dF(t) = -\kappa(F(t) - \mu(t)) dt + \sigma dW(t) + \sum_j \eta_j \delta(t - t_j).$$

Measured Ledger (continuous-time idealization):

$$\hat{L}(t) = \int_0^t HCl(\tau) d\tau.$$

Measured Ledger (discrete-time approximation):

$$\hat{L}(t_n) = \sum_i HCl(t_i) \cdot \Delta t_i, \text{ where } \Delta t_i = t_i - t_{i-1}.$$

Missingness Model (illustrative):

$$\Pr(M_t^{(i)} = 1 | x_t^{(i)}, W_t^{(i)}) = \text{logit}^{-1}(\alpha + \gamma x_t^{(i)} + c^\top W_t^{(i)}).$$

Missingness (Selection) Model (simple illustrative form):

$$\Pr(M_t = 1 | F_t) = \text{logit}^{-1}(\alpha_0 + \alpha_1 F_t).$$

Neural Control Regression (example):

$$y_{ROI}(t) = \beta_0 + \beta_1 \Phi(u_t; L(t), H_t) + \beta_2 U(t) + \beta_3 C(t) + \beta_4 \text{Arousal}(t) + \varepsilon(t).$$

Neural Control Regression (example; alternate form):

$$Y(t) = \beta_0 + \beta_1 U + \beta_2 \text{Conf} + \beta_3 \text{Arousal} + \beta_4 \Phi + \varepsilon.$$

Neutrality Band:

$L(T) \in [-K, K]$, where K is the preregistered neutral-band half-width.

Neutrality Gate (Bayesian):

$$p_{\text{neutral}}(T) \geq 1 - \varepsilon.$$

Neutrality Convergence Statement (idealized claim statement):

$$\lim_{\{t \rightarrow T^-\}} L(t) = 0 \text{ (within tolerance } K).$$

Operational Stopping Time (unity-based; simple form):

$$T = \inf\{ t \geq 0 : \text{UnityIndex}(t) < \theta \}.$$

Optional Stopping (one standard statement):

If $M(t)$ is a martingale and τ is an admissible stopping time, then $E[M(\tau)] = E[M(0)]$.

Ornstein–Uhlenbeck Affect Dynamics:

$$dF(t) = -\kappa(F(t) - \mu(t)) dt + \sigma dW(t).$$

Population-Level Variance Bound (example framing):

$\text{Var}[L(T)] \rightarrow c$ within a preregistered bound,

with terminal compression assessed via decline in variance near the end (e.g.,

$$(d/dt)\text{Var}[L(t)] < 0 \text{ in the endgame window}).$$

Population Shadow Price (heuristic):

$$\Lambda_t \equiv \partial/\partial R \Pr(\bigcap_i L_i(T_i) \in [-K, K]).$$

Posterior Draws → Terminal Ledger:

$$\hat{L}^{\{(m)\}}(T) = \sum_t \Delta t_t \cdot x_t^{\{(m)\}}.$$

Posterior Gate (terminal neutrality):

$$\Pr(|L(T)| \leq K | \text{data}) \geq 1 - \varepsilon.$$

Posterior Neutrality Probability:

$$p_{\text{neutral}}(T) = P(L(T) \in [-K, K] | \text{data}) \approx (1/M) \sum_m 1[\hat{L}^{\{(m)\}}(T) \in [-K, K]].$$

Posterior Neutrality Criterion (Bayesian acceptance criterion; example):

$$\Pr(|L(T)| < K | \text{data}) > 0.95.$$

Queueing Congestion Proxy (M/M/1 example):

$$\rho = \lambda_{\text{arr}}/\mu, \text{ and } E[W_q] = \rho / [\mu(1 - \rho)].$$

Queueing Wait-Time Heuristic (M/M/1; same expression, alternate notation):

$$E[W_q] = \rho / [\mu(1-\rho)], \text{ where } \rho = \lambda_{\text{arr}}/\mu.$$

QS-Residual Signal Model (example):

$$A(t) = f(\text{recent history}) + \beta \cdot \Phi(L(t), H_t) + \varepsilon.$$

Repair/Relief Choice Model (example):

$$\text{logit } P(\text{choose repair/relief at } t) = \alpha + \beta_H g(H_t) + \beta_\Phi \Phi_t + \beta_{\{g \times \Phi\}} \cdot (g(H_t) \times \Phi_t), \\ \text{with } g(H_t) = H_t^{-1}.$$

Regression Discontinuity (threshold design; example):

$$Y_i = \tau \cdot 1[X_i \geq c] + f(X_i - c) + v_i.$$

Resource Penalty:

$$\text{Penalty}(u, t) = \sum_r \lambda_{\{r\}}(t) \cdot \Delta_r(u).$$

Resource-Gated Action Menu (resource-aware admissibility):

$$\mathcal{A}(t | R) = \{ u \in \mathcal{S}(t) : \Pr\{ L(T) \in [-K, K] | \pi_u, L(t), H_t, R(t) \} \geq 1 - \varepsilon \}.$$

Sleep-Stage Variance Modulation (illustrative):

$$\sigma^2_{\{\eta, t\}} = \sigma_0^2 \exp(-\kappa_{\text{SWS}} \cdot 1_{\text{SWS}}(t) + \kappa_{\text{REM}} \cdot 1_{\text{REM}}(t)).$$

Softmax Action Weight (minimal):

$$\omega(u; t) \propto \exp[\beta(H_t) \cdot \Phi(u; L(t), H_t)].$$

Softmax Choice Probability (expanded example):

$$\Pr(u_t = u) = \frac{\exp[\theta_0 + \theta_u \cdot U_{\text{standard}}(u) + \theta_C \cdot C(u) + \theta_A \cdot \text{Arousal} + \theta_\Phi \cdot \Phi(u; L(t), H_t)]}{\sum_{v \in \mathcal{S}(t)} \exp[\theta_0 + \theta_u \cdot U_{\text{standard}}(v) + \theta_C \cdot C(v) + \theta_A \cdot \text{Arousal} + \theta_\Phi \cdot \Phi(v; L(t), H_t)]}.$$

State-Change Valence:

$$V(t) = F(t) = -W \cdot dS(t)/dt.$$

State-Space Latent Affect (illustrative):

$$x_{t-1}^{(i)} = \phi x_{t-1}^{(i)} + b^T Z_t^{(i)} + u_i + \eta_t^{(i)}, \\ \text{and } y_{c,t}^{(i)} = a_c x_{t-1}^{(i)} + d_c Z_t^{(i)} + \varepsilon_{c,t}^{(i)}.$$

State-Space (latent affect evolution; linear-Gaussian example):

$$F_t = F_{t-1} + \alpha^T u_t + \eta_t, \eta_t \sim \text{Normal}(0, Q).$$

State-Space Observation (vector form; linear-Gaussian example):

$$\tilde{y}_t = \Lambda F_t + \Gamma Z_t + \varepsilon_t, \varepsilon_t \sim \text{Normal}(0, \Psi).$$

Stochastic Ledger Dynamics (Itô form):

$$dL_t = \mu(t) dt + \sigma(t) dW_t.$$

Terminal Neutrality Band (alternate statement):

$L(T) \in [-K, +K]$ (equivalently, $|L(T)| < K$).

Terminal Time as Stopping Rule (unity-based; extended):

$T = \inf\{ t \geq 0 : \text{UnityIndex}(t) < \theta \text{ AND no return within } \tau_{\text{pause}} \}$.

Terminal Window Mean:

$$HCl_{-\tau}(T) = (1/\tau) \int_{T-\tau}^T HCl(t) dt.$$

Terminal-Window Variance Metric:

$$\sigma^2_{-\tau}(T) = \text{Var}[HCl(t)] \text{ for } t \in (T-\tau, T];$$

a common summary is $\text{Var}_{\text{end}} / \text{Var}_{\text{base}}$.

Transformed Process (martingale transform; same as above, alternate label):

$$M(t) = L(t) + \lambda \int_0^t \Psi(L(s), H_s) ds.$$

Under standard regularity conditions, $dM(t) = \sigma dW(t)$.

Two-State Change Link to HCl (illustrative):

$$\Delta HCl_t = \alpha_F \cdot \Delta F_t + \alpha_G \cdot \Delta G_t + \varepsilon_t.$$

Uncertainty Propagation (discrete integral; generic):

$$\text{Var}[L(T)] = \sum_m \sum_n \text{Cov}(\hat{F}(t_m), \hat{F}(t_n)) \cdot \Delta t_m \cdot \Delta t_n.$$

Variance Ratio Metric (end-of-life):

$$\rho_i = \sigma^2_{\{\tau, i\}(T)} / \sigma^2_{\{\text{base}, i\}}.$$

Citations and References

- Adams, J. S. (1963). Toward an understanding of inequity. *Journal of Abnormal and Social Psychology*, 67, 422–436.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *Proceedings of the 2nd International Symposium on Information Theory* (pp. 267–281). Akadémiai Kiadó.
- Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex: one decade on. *Trends in Cognitive Sciences*, 8(4), 170–177.
- Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books.
- Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.
- Barrett, L. F. (2006). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review*, 10(1), 20–46.
- Berridge, K. C., and Kringelbach, M. L. (2015). Pleasure systems in the brain. *Neuron*, 86(3), 646–664.
- Brickman, P., Coates, D., and Janoff-Bulman, R. (1978). Lottery winners and accident survivors: Is happiness relative? *Journal of Personality and Social Psychology*, 36(8), 917–927.
- Brosnan, S. F., & de Waal, F. B. M. (2003). Monkeys reject unequal pay. *Nature*, 425, 297–299.
- Brosnan, S. F., & De Waal, F. B. (2014). Evolution of Responses to (Un)fairness. *Science*, 346(6207), 1251776.
- Cabanac, M. (1971). Physiological role of pleasure. *Science*, 173(4002), 1103–1107.
- Cannon, W. B. (1932). *The Wisdom of the Body*. W. W. Norton and Company.
- Carr, M.-F., Jadhav, S. P., & Frank, L. M. (2011). Hippocampal replay in the awake state: sequential experience replays during awake slow-wave sleep in a circular track. *Neuron*, 67(2), 233–243.
- Carstensen, L. L. (1999). Taking time seriously: a theory of socioemotional selectivity. *American Psychologist*, 54, 165–181.

- Cartwright et al. (1984): Cartwright, R. D., Lloyd, S., Knight, S., & Trenholme, I. (1984). *Broken dreams: A study of the effects of divorce and depression on dream content*. Psychiatry, 47(3), 251–259.
- Cartwright et al. (2006): Cartwright, R., Agargun, M. Y., Kirkby, J., & Friedman, J. K. (2006). *Relation of dreams to waking concerns*. Psychiatry Research, 141(3), 261–270.
- Carver, C. S., and Scheier, M. F. (1990). Origins and functions of positive and negative affect: A control-process view. *Psychological Review*, 97(1), 19–35.
- Charles, S. T., Mather, M., & Carstensen, L. L. (2003). Aging and Emotional Memory: The Forgetting of Sad Faces. *Journal of Experimental Psychology: General*, 132(2), 310–324.
- Clark, A. (2015). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.
- Critchley, H. D., Wiens, S., Rotshtein, P., Öhman, A., and Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience*, 7(2), 189–195.
- Damasio, A. R. (1996). *Descartes' Error: Emotion, Reason, and the Human Brain*. Penguin Books.
- Diener, E., Lucas, R. E., and Scollon, C. N. (2006). Beyond the hedonic treadmill: Revising the adaptation theory of well-being. *American Psychologist*, 61(4), 305–314.
- Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Econometrica*, 67(4), 1241–1283.
- Fields, H. L. (2004). State-dependent opioid control of pain. *Nature Reviews Neuroscience*, 5(7), 565–575.
- Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology*, 65(1), 45–55.
- Friston, K. J. (2010). The free-energy principle: a unified brain theory?. *Nature Reviews Neuroscience*, 11(2), 127–138.
- Goldstein, A. N., and Walker, M. P. (2014). The role of sleep in emotional brain function. *Annual Review of Clinical Psychology*, 10, 679–708.
- Gould, S. J., and Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society B*, 205(1161), 581–598.

- Gross, J. J., and Jazaieri, H. (2014). Emotion, emotion regulation, and psychopathology: An affective science perspective. *Clinical Psychological Science*, 2(4), 387–401.
- Hamilton, W. D. (1964). The genetical evolution of social behaviour I. *Journal of Theoretical Biology*, 7(1), 1–16.
- Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press.
- Joffily, M., & Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS Computational Biology*, 9(6): e1003094.
- Juechems, K., Balaguer, J., Herce Castañoñ, S., Ruz, M., O'Reilly, J. X., & Summerfield, C. (2019). A network for computing value equilibrium in the human medial prefrontal cortex. *Neuron*, 101(5), 977–987.e3.
- Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, 4(6), 401–405.
- Keramati, M., and Gutkin, B. S. (2014). Homeostatic reinforcement learning for integrating reward collection and physiological stability. *eLife*, 3, e04811.
- Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, 314(5800), 829–832.
- Koob, G. F., & Le Moal, M. (2008). Addiction and the brain antireward system. *Annual Review of Psychology*, 59, 29–53.
- Leknes, S., Brooks, J. C., Wiech, K., & Tracey, I. (2008). Pain relief as an opponent process: a psychophysical investigation. *European Journal of Neuroscience*, 28(4), 794–801.
- Lerner, M. J. (1980). *The Belief in a Just World: A Fundamental Delusion*. Plenum Press.
- Lyubomirsky (2011): Lyubomirsky, S. (2011). *Hedonic adaptation to positive and negative experiences* (pp. 200–224). In S. Folkman (Ed.), *Oxford handbook of stress, health, and coping*. New York: Oxford University Press.
- Mather, M., and Carstensen, L. L. (2005). Aging and motivated cognition: The positivity effect in attention and memory. *Trends in Cognitive Sciences*, 9(10), 496–502.
- McEwen, B. S., & Stellar, E. (1993). Stress and the individual: mechanisms leading to disease. *Archives of Internal Medicine*, 153(18), 2093–2101.
- Nagel, T. (1979). *Mortal Questions* (ch. “Moral Luck”). Cambridge University Press.
- Nesse, R. M., and Williams, G. C. (1994). *Why We Get Sick: The New Science of Darwinian Medicine*. Times Books.

- Nishida, M., Pearsall, J., Buckner, R. L., & Walker, M. P. (2009). REM sleep, prefrontal theta, and the consolidation of human emotional memory. *Cerebral Cortex*, 19(5), 1158–1166.
- Noether, E. (1918). Invariant variation problems. *Nachrichten der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 235–257.
- Ochsner, K. N., & Gross, J. J. (2005). The cognitive control of emotion. *Trends in Cognitive Sciences*, 9(5), 242–249.
- Panksepp, J. (2005). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press.
- Pascal, R., & Pross, A. (2015). Stability and its manifestation in the chemical and biological worlds. *Chemical Communications*, 51(70), 16160–16165.
- Perogamvros, L., Schwartz, S., and Ruby, P. (2023). *On dreams regulating emotions by placing threats in safe contexts*. Consciousness and Cognition, 115, 103534.
- Popper, K. (1959). *The Logic of Scientific Discovery*. Routledge.
- Pross, A. (2012). *What is Life? How Chemistry Becomes Biology*. Oxford University Press.
- Pross, A., & Pascal, R. (2017). How and why kinetically stable systems form far from equilibrium. *Philosophical Transactions of the Royal Society B*, 372(1726), 20160431.
- Ramsay, D. S., and Woods, S. C. (2014). Clarifying the roles of homeostasis and allostasis in physiological regulation. *Psychological Review*, 121(2), 225–247.
- Saurat, M.-T., Agbakou, M., Attigui, P., Golmard, J.-L., and Arnulf, I. (2011). *Walking dreams in congenital and acquired paraplegia*. Consciousness and Cognition, 20(4), 1425–1432
- Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17(11), 565–573.
- Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306), 1593–1599.
- Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, 303(5661), 1157–1162.

- Solomon, R. L., and Corbit, J. D. (1974). An opponent-process theory of motivation: I. Temporal dynamics of affect. *Psychological Review*, 81(2), 119–145.
- Solomon (1980): Solomon, R. L. (1980). *The opponent-process theory of acquired motivation: The costs of pleasure and the benefits of pain*. American Psychologist, 35(8), 691–712
- Stafford, E. (2000). *Worshipping Virtues: Personification and the Divine in Ancient Greece*. London: Duckworth.
- Sterling, P. (2012). Allostasis: A model of predictive regulation. *Physiology and Behavior*, 106(1), 5–15.
- Taquet, M., Quoidbach, J., de Montjoye, Y.-A., Desseilles, M., and Gross, J. J. (2020). Mood homeostasis, low mood, and history of depression. *JAMA Psychiatry*, 77(9), 944–951.
- Tedeschi, R. G., and Calhoun, L. G. (2004). Posttraumatic growth: Conceptual foundations and empirical evidence. *Psychological Inquiry*, 15(1), 1–18.
- Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *Biological Bulletin*, 215, 216–242.
- Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46(1), 35–57.
- Vaillant, G. E. (2012). *Triumphs of Experience: The Men of the Harvard Grant Study*. Harvard University Press.
- Van der Helm, E., and Walker, M. P. (2009). Overnight therapy? The role of sleep in emotional brain processing. *Psychological Bulletin*, 135(5), 731–748.
- Wagner, N., & Pross, A. (2011). The nature of stability in replicating systems. *Entropy*, 13(2), 518–527.
- Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C.-W., & Kross, E. (2013). An fMRI-based neurologic signature of physical pain. *New England Journal of Medicine*, 368, 1388–1397.
- Wolf, S. (1982). Moral saints. *Journal of Philosophy*, 79(8), 419–439.
- Yoo, S.-S., Gujar, N., Hu, P., Jolesz, F. A., and Walker, M. P. (2007). The human emotional brain without sleep—A prefrontal amygdala disconnect. *Current Biology*, 17(20), R877–R878.

