The Redemption Optimization

Why Evil Exists in a Divine Blueprint for Ultimate Good

Sergiu Margan

September 8, 2025

Contents

Abstract

The Redemption Optimization proposes an event–valued framework for reasoning about evil, freedom, and the highest good, and then operationalises it for ethics, AI safety, and public governance. The core move is to evaluate *what actually happens in history*—enacted protections (ΔL), exercised freedoms (ΔF), and realized acts of mercy and justice (M, J^{v})—rather than capacities, plans, or near–misses. Formally, we study an objective

$$\mathcal{J} = \sum_{t \geq 0} \left(\alpha \, \Delta L_t + \gamma \, \Delta F_t + \mu \, M_t + \nu \, J_t^{v} \right) - \sum_{t \geq 0} \left(\beta \, H_t + \kappa \, R_t \right),$$

subject to bounded cumulative harm and non–coercion. A *rejection* R_t (counted wrongdoing) occurs iff a rights/standards breach or harm above threshold is present and culpability is adjudicated.

Two main results structure the book. First, the **Two–Branch Fork** ("Margan's Optimization Paradox"): if redemption–goods are valued ($\mu + \nu > 0$), then along any history with no counted rejection ($\sum_t R_t = 0$) the event–valued redemption credits are necessarily zero. Thus the evil–free branch is coherent only when redemption–goods

are not valued in the event sense; otherwise a minimal realized refusal is required to ground mercy/justice in actuality. Second, **Minimal–Trigger Optimality**: with typal/concave kind–credits and positive per–event costs, at most *one* gated rejection per class before closure maximizes $\mathcal{J}$; repetition adds cost without kind–level value.

We formalise **confirmation** (S^*) as *structural hazard removal*: a three–stage operator—cohort replay, stress drills under stronger incentives, and a bounded monitored–live window—certifies that the prior causal route is closed (HZ = 0) while alternate possibilities remain live (non–coercion). Near–misses and drills support confirmation but do not mint redemption credit.

Applications span theology and practice. Theologically, the framework reframes classic theodicies and maps the arc Permission $\rightarrow$ Rectification $\rightarrow$ Confirmation to the Biblical narrative, clarifying where and why evil is permitted and how eternal freedom coexists with secured good. Practically, event–valued ledgers, minimal–trigger doctrine, and confirmation tests are instantiated for AI systems and safety engineering, healthcare (sentinel events), elections, disaster response, finance, supply chains, policing/justice, and other civic domains.

The volume includes rigorous assumptions and proofs, domain playbooks, and replication materials (schemas, metrics, ablations, falsification tests). A companion repository enables figure and table regeneration from configuration files. Overall, the book offers a single evaluator that is simultaneously faithful to Scripture's moral grammar and fit for technical audit: *count events, fix routes, prove closure, keep freedom live*.

Keywords: event–valued ethics; theodicy; freedom and confirmation; AI safety; governance; minimal–trigger optimality; structural closure.

How to Read This Book

Notation

What this section is. A short guide for different readers—general, theological, philosophical, technical, and policy—showing the fastest routes through the material, how the symbols/boxes work, and where to find proofs, playbooks, and replication artefacts. Keep this open while you read.

Who you are & your fastest route

General reader (no math required). Start with the *Abstract* and *Preface*, then read the narrative chapters that explain the core idea without proofs:

Chs. 1–3 ⇒ Ch. 12 (The Optimal Arc) ⇒ Ch. 13–16 (story & synthesis).

Dip into the *Casebook* to see how it plays out in real life (Chs. 17–36). Ignore equations on a first pass; every formula has a plain-language paraphrase nearby.

Theology & ministry. Read Chs. 1–3 (problem of evil), then Chs. 12–16 (permission → rectification → confirmation) and selected applications close to pastoral practice (e.g., Chs. 29–33). Appendix B gives a symbol crosswalk from theological terms to the formal ledger.

Philosophy (ethics, decision theory, metaphysics). Skim Ch. 1 for the framing, then read Chs. 7–11 (formal core) plus App. A (assumptions, evaluators, proofs). Pay attention to the fork (MOP), the gate for R_t, and the difference between event-valued $\mathcal{J}$ and capacity-valued $\mathcal{J}^\dagger$.

AI & safety engineering. Start with Ch. 9–11 (minimal-trigger doctrine and confirmation), then jump to domain instantiations (AI chapter and safety-critical chapters in the Casebook) and Ch. 41 (methods and replication). Use Appendix G to regenerate the figures/tables.

Law, policy, governance. Read Chs. 12–14 for the doctrine and S^* (confirmation), then your sector chapters in the Casebook. Appendices E–F contain printable scope locks, dockets, confirmation plans, and one-page posters you can adopt directly.

Map of the book (at a glance)

- **Chs. 1–6** — Problem parameters, prior art, and methods.
- **Chs. 7–11** — *Formal core:* evaluator $\mathcal{J}$, MOP (Two-Branch Fork), minimal-trigger optimality, and S^* (confirmation).

- **Chs. 12–16** — *The optimal arc* and theological synthesis.
- **Chs. 17–36** — *Casebook:* sector playbooks with triggers, dockets, and tests.
- **Chs. 37–40** — *Pillars:* pillar domains (elections, disaster response, finance, supply chain) in depth.
- **Ch. 41** — *Methods:* datasets/generators, metrics, tests, ablations, falsifiers, and full replication protocol.
- **Apps. A–G** — Assumptions & proofs; glossary; audit checklists; worked examples; forms; posters; replication index.

Reading paths (choose one)

Two-hour tour (no equations)

Abstract → Ch. 1 (Problem) → Ch. 12 (Arc) → one Casebook chapter you care about → Conclusion. Keep Appendix B open for symbols.

Technical backbone (half-day)

Ch. 7 (Evaluator) → Ch. 8–11 (Fork, Minimal-trigger, Confirmation) → App. A (assumptions/proofs) → Ch. 41 (methods) → App. G (commands) → your domain chapter.

Policy implementer (playbooks first)

Ch. 12–14 (doctrine) → your Casebook chapter(s) → Apps. E–F (forms/posters). Return to Chs. 7–11 for the rationale if challenged.

How the boxes and labels work

- **Assumption Boxes (A1–A8):** grey boxes that state guardrails (event-valued accounting, non–coercion, bounded harm, honesty locks, etc.). Each theorem cites the exact A# it uses (see App. C).
- **Notation Boxes:** quick reminders of symbols used in the surrounding section.
- **Definition / Lemma / Theorem / Corollary:** formal objects. Proofs or proof sketches follow or are expanded in App. D.
- **Practice Rules & Checklists:** operational versions of the theory; many are reproduced as printable posters in App. F.

Minimal math primer (for non-specialists)

- **Event-valued vs capacity-valued:** does credit attach to *what actually happened* ($\mathcal{J}$) or to *what could have happened* ($\mathcal{J}^{\dagger}$)?
- **Rejection (R_t):** counted only if there is a breach or harm over a floor *and* culpability is present (A3).

- **Confirmation (S^*):** a three-stage structural proof that the hazardous route is closed while freedom remains live (A7).
- Full symbol list: Appendix B.

What to skip on a first pass

If you only want the idea: skip formal proofs in Chs. 7–11 and read the plain text around them; proofs are collected and expanded in Apps. A and D. In the Casebook, read the "Trigger → Docket → Confirmation" subsections first.

Replication and artifacts

Figures and tables are reproducible from a public repository. Use Ch. 41 for methods and Appendix G for exact commands and checksums. If you are reviewing, see App. C for audit checklists.

Claims discipline (how to read assertions)

Statements are *conditional*: "Given A1–A4, Theorem X follows." When we switch evaluators (from $\mathcal{J}$ to $\mathcal{J}^\dagger$), we say so explicitly. Domain chapters state thresholds and pass/fail criteria up front.

A final suggestion

Read with a pencil. When you see M (mercy) or J^v (justice) credited, ask: *What event minted it?* When you see $S^* = 1$, ask: *Where are*

the cohort replay, stress drills, and monitored-live evidence? Those two questions will carry you through the entire book. “‘0

Research Questions

Notation

Purpose. This page states the book's research questions (RQs), grouped by theory, methods, domains, and policy. Each RQ is framed so that it is (i) auditable under our assumptions A1–A8 (Apps. A, C), (ii) paired with concrete tests or artefacts (Ch. 41; App. G), and (iii) falsifiable by specific counterevidence.

A. Formal Core (theorems and comparators)

RQ-A1 Event-valued necessity (Fork / MOP). Under A1–A3, A6, does valuing realised redemption-goods ($\mu + \nu > 0$) imply that any evil-free history ($\sum_t R_t = 0$) earns zero redemption credit? *Test:* reproduce Thm. (Two-Branch Fork) on synthetic ledgers and show that, under $\mathcal{J}$, M, J^{ν} remain zero when $R \equiv 0$; contrast with $\mathcal{J}^{\dagger}$ (App. A, D).

RQ-A2 Minimal-trigger optimality. With typal/concave kind-credits and positive per-event costs (A4–A5), is $n_c = 1$ (one gated

rejection per class before closure) optimal among policies that close the route? *Test:* ablations over ϕ_c and (β, κ); demonstrate domination of $n_c > 1$ (App. D).

RQ-A3 Non-coercive confirmation. Can S^* be realised as structural hazard removal (A2, A7–A8) while alternate possibilities remain live? *Test:* graph-cut formalisation + three-stage plan with pre-registered pass criteria (App. D).

RQ-A4 Adjudicating evaluators. In practice, when does event-valued $\mathcal{J}$ outperform capacity-valued $\mathcal{J}^\dagger$ in tracking real outcomes? *Test:* paired evaluations on the same datasets; report where the evaluators diverge and which predicts recurrences better (Ch. 41).

B. Measurement & Methods

RQ-B1 Measuring ΔL and ΔF. What robust, privacy-preserving proxies estimate delivered protections and exercised freedoms with acceptable inter-rater reliability? *Test:* codebooks, IRR statistics, and sensitivity analyses (Ch. 41).

RQ-B2 Culpability gate reliability. How consistent is C_t adjudication across raters and domains? *Test:* blinded vignettes; confusion matrices; adjudicator guidelines (App. C).

RQ-B3 Thresholds and harm floors. How should $h_{\min}$ and $H_{\max}$ be set to balance sensitivity and over-blocking? *Test:* ROC-style sweeps; report two-sided harm $(H^{\text{miss}}, H^{\text{over}})$ where applicable (Ch. 41).

RQ-B4 Closure latency. What are typical T_c^* (time to confirmation) distributions by domain, and which interventions reduce them? *Test:* cohort comparisons before/after adopting the doctrine; SPC charts (Ch. 41).

C. Theology & Philosophy

RQ-C1 Narrative mapping. Does the arc Permission → Rectification → Confirmation faithfully reflect Scripture's moral grammar without special pleading? *Test:* exegesis across representative passages; show where mercy/justice are *events* (not merely capacities).

RQ-C2 Modal vs event-valued talk. Where do classic theodicies implicitly switch to capacity-valued reasoning, and can that be made explicit and fair? *Test:* literature review (Chs. 4–6) with formal side-by-side examples.

RQ-C3 Freedom and confirmation. Is "live AP with hazard removed" conceptually coherent and pastorally acceptable? *Test:* philosophical analysis + case analogies (Chs. 12–16).

D. Domains (Casebook and Pillars)

RQ-D1 AI/Platforms. Does Trigger → Docket → S^* reduce policy-violations and harmful outputs while preserving useful capability? *Metrics:* violation rate, time-to-fix, tool-gating efficacy, user value retained.

RQ-D2 Healthcare. Do sentinel-event closures reduce recurrence without coercive side-effects (e.g., harmful throughput throttling)? *Metrics:* recurrence by class, T_c^*, safety vs flow.

RQ-D3 Elections. Do scope-locked closures (e.g., `TAB-ERR`, `ACCESS-FAIL`) cut repeats across cycles? *Metrics:* RLA results, accessibility compliance, error rates.

RQ-D4 Disaster/EM. Do warning/evacuation closures pass stress drills and monitored-live checks? *Metrics:* alert reach/time, evac completion, drill scores.

RQ-D5 Finance. Can sanctions/fraud classes achieve higher recall with bounded over-blocking? *Metrics:* recall/precision, victim loss, lawful-user denial minutes.

RQ-D6 Supply chain/food. Do allergen/pathogen closures cut exposure-days and improve recall reach/time? *Metrics:* exposure, recall coverage/latency, waste (over-recall).

E. Policy & Governance

RQ-E1 Minimal-trigger doctrine in law. When encoded in policy, does "one counted wound per class $\rightarrow$ closure" reduce recurrence faster than training-only regimes? *Test:* difference-in-differences across jurisdictions/units adopting the doctrine.

RQ-E2 Honesty locks. Do anti-suppression measures (random checks, whistleblowing) lower false closure claims without chilling le-

gitimate activity? *Test:* pre/post incident reporting rates; audit outcomes.

RQ-E3 **Equity.** Do $\Delta L, \Delta F$ distributions improve across protected groups under the framework? *Test:* slice metrics with privacy-safe reporting (Ch. 41).

F. Falsifiers (what would disconfirm claims?)

F1 A reproducible policy or system where $n_c > 1$ strictly improves $\mathcal{J}$ *while* A1–A5 hold (i.e., typal/concave credits and positive costs are satisfied).

F2 A confirmed closure ($S^* = 1$) that later repeats via the *same* causal vector without scope tampering (violates A8) across multiple audits.

F3 A non-coercive alternative to S^* that achieves strictly better outcomes on both harm reduction and freedom metrics with the same evidence rigor.

F4 Robust real-world cases where event-valued $\mathcal{J}$ systematically underpredicts recurrences compared to $\mathcal{J}^\dagger$ across domains.

F5 Evidence that $\Delta L/\Delta F$ cannot be measured with adequate reliability despite clear codebooks and training.

G. Mapping RQs to Chapters & Artefacts (at a glance)

ID	Short name	Primary chapters	Tests / Artefacts
RQ-A1	Fork / MOP	Chs. 7–9; App. A	Synthetic ledgers; $\mathcal{J}$ vs $\mathcal{J}^{\dagger}$ plots
RQ-A2	Minimal-trigger	Chs. 9–11; Apps. A,D	Ablations; typal ϕ_c table; weight sweeps
RQ-A3	Non-coercive S^*	Ch. 11–12; Apps. A,D	Graph cut + 3-stage plan; pass criteria
RQ-A4	Evaluator adjudication	Ch. 41; App. G	Paired metrics; recurrence prediction gap
RQ-B1	$\Delta L, \Delta F$ measurement	Ch. 41	IRR stats; privacy-preserving proxies
RQ-B2	Culpability gate	Ch. 41; App. C	Vignette study; adjudicator guidelines
RQ-B3	Threshold tuning	Ch. 41	ROC-style

H. Success Criteria (summary)

- **Theoretical:** proofs replicate; assumptions usage is explicit; counterexamples locate which A# fails.
- **Empirical:** figures/tables regenerate from configs; seeds and checksums match (App. G).
- **Domain:** recurrence falls, T_c^* shortens, harm budgets respected, $\Delta L, \Delta F$ improve without coercion.
- **Governance:** honesty locks raise integrity of closure without suppressing legitimate activity; scope-locks prevent relabelled repeats.

Contributions

Notation

What this book contributes. Below we list the principal theoretical, methodological, and practical contributions, the artifacts that make them reproducible, and what we *do not* claim. Cross-references point to appendices for formal definitions and audit trails.

1. Core Theoretical Contributions

1. **Event-valued evaluator (single objective).** We formalise an objective that credits only *realised* goods and enacted remedies:

$$\mathcal{J} = \sum_{t \geq 0} \left(\alpha \, \Delta L_t + \gamma \, \Delta F_t + \mu \, M_t + \nu \, J_t^{v} \right) - \sum_{t \geq 0} \left(\beta \, H_t + \kappa \, R_t \right),$$

   with bounded cumulative harm and non–coercion guardrails (definitions in App. A, glossary in App. B).

2. **Two-Branch Fork (Margan's Optimization Paradox, MOP).** If redemption-goods are valued ($\mu + \nu > 0$), an evil-free history

($\sum_t R_t = 0$) necessarily earns zero event-valued redemption credit. Therefore either (i) a minimal realised refusal occurs (grounding mercy/justice), or (ii) redemption-goods are valued only modally/capacitively (App. A).

3. **Minimal-Trigger Optimality.** With typal/concave kind-credits and positive per-event costs, at most *one* gated rejection per class before closure maximises $\mathcal{J}$; repetition adds cost without kind-level value (App. A, proofs expanded in App. D).

4. **Confirmation as structural hazard removal.** We formalise S^* as a three-stage operator—cohort replay, stress drills, monitored-live—that certifies route closure (HZ = 0) while alternate possibilities remain live (non–coercion). Graph view and soundness bounds are provided (Apps. A, D).

5. **Capacity-valued comparator.** We separate an alternative evaluator $\mathcal{J}^\dagger$ that credits capacities to show precisely where modal talk diverges from event-valued ethics (Apps. A, D).

2. Methods & Reproducibility Contributions

1. **Event schemas & ledgers.** Minimal, domain-agnostic schemas for recording $R, B, C, H, \Delta L, \Delta F, M, J^\nu$, HZ, S^*; per-domain extensions (Ch. 41, App. B).

2. **Metrics & dashboards.** Closure latency T_c^*, recurrence by class, harm budgets E_{tot}, and distributional reporting for $\Delta L, \Delta F$ with equity slices (Ch. 41).

3. **Ablations & falsification tests.** Weight sweeps, typal/concave vs linear credits, harm floors, culpability gate toggles; explicit disconfirmers (Ch. 41, App. C).

4. **Replication artefacts.** A command-indexed appendix mapping figures/tables to exact notebooks/CLI calls plus checksum manifests (App. G).

3. Practical & Policy Contributions

1. **Minimal-trigger doctrine for governance.** A per-class rule: first validated $R^{(c)} = 1 \Rightarrow$ open remediation docket $\Rightarrow$ enact mercy/justice $\Rightarrow$ confirm and scope-lock; repetition earns no kind-level bonus.

2. **Sector playbooks.** Instantiations for AI & platforms, elections, disaster response, healthcare, policing/justice, finance, supply chain/food safety, energy/utilities, and others. Each chapter provides triggers, dockets, confirmation tests, anti-gaming checks, and dashboards.

3. **Closure packs & forms.** Ready-to-use scope locks, dockets, confirmation plans, evidence manifests, and public summaries (App. E); printable posters and decision cards for teams (App. F).

4. What Is *Not* Claimed

- Not a proof of metaphysical necessity beyond the stated assumptions; all theorems are conditional and auditable (Apps. A,

C).

- Not a replacement for domain-specific safety standards; rather, a unifying evaluator and confirmation doctrine that can sit atop them.

- Not a guarantee against all recurrence; scope-locked closure attests structural removal for a defined causal route, with honest monitoring bounds.

5. Reusability & Artifacts

- **Code/data.** A companion repository regenerates all figures/tables from configuration (*weights.yaml*, *thresholds.yaml*, *seeds.yaml*); command index in App. G. (DOI to be added at release.)

- **Licensing.** Book text (standard copyright). Example code and templates released under a permissive licence suitable for academic/operational reuse (to be specified in the repo).

- **Versioning.** Tagged release `v1.0-book` corresponds to this edition; subsequent editions will update App. G.

6. Impact Summary (Discipline-specific)

Theology & Philosophy	A precise, event-valued reframing of theodicy; MOP clarifies why redemption-goods require realised refusal under stated assumptions.
AI & Safety Engineering	A single evaluator for before/after tests, tool-gating, and structural confirmation; near-misses become drills that support (but do not counterfeit) closure.
Law & Governance	Minimal-trigger rule operationalised into dockets, artefacts, and public summaries; scope locks prevent relabelling repeats.
Healthcare, Elections, Supply Chain, Finance	Domain-specific triggers, tests, and dashboards to align incentives with event-level outcomes.

7. Open Problems

1. **Measuring $\Delta L, \Delta F$ at scale.** Robust, privacy-preserving proxies that still retain auditability.

2. **Causal-vector discovery.** Automating the formation of class scopes c and detecting when a recurrence is "same route" vs a genuinely new c'.

3. **Game-theory of honesty locks.** Incentive-compatible mechanisms that deter suppression without chilling legitimate activity.

4. **Cross-evaluator adjudication.** Empirically adjudicating $\mathcal{J}$ vs $\mathcal{J}^{\dagger}$ in domains where capacities are tempting to count.

One-line contribution. A single, auditable evaluator—together with a minimal-trigger doctrine and structural confirmation—linking theological insight to technical practice: *count events, fix routes, prove closure, keep freedom live*.

Publisher's Note

Edition & Imprint

Title: *The Redemption Optimization: Why Evil Exists in a Divine Blueprint for Ultimate Good*
Author: Sergiu Margan
Edition: First edition (2025)
Imprint: Independent Imprint (TBD), TBD
Print: Printed in the United Kingdom on acid-free paper.

Copyright & Rights

Open materials. Selected figures, tables, and templates are reproducible from the companion repository under a permissive licence; see Ch. 41 and App. G for commands and licence details.

A full replication pack, including simulations and proofs for the Margan Optimization Paradox, is permanently archived on Zenodo at DOI: 10.5281/zenodo.17079986. Reuse of those *code-generated artefacts* must follow the repository licence; reuse of book prose and page layouts requires permission from the publisher.

Disclaimers

This volume presents research and opinion intended for scholarly and professional audiences. It does not constitute legal, medical, clinical, investment, or safety certification advice. The publisher and author have made every effort to ensure the accuracy of information at the time of publication and assume no liability for errors or omissions. Examples and case studies are anonymised, synthetic, or public; where partner procedures informed a template, only redacted/aggregate artefacts are reproduced.

Trademarks & Terms

"**Margan's Optimization Paradox (MOP)**" and "**Margan's Optimization Framework**" are descriptive names used in this work. Trademark symbols (^{TM}/$^{®}$) are omitted in running text for readability. All other product and company names are trademarks of their respective holders; use herein does not imply affiliation or endorsement.

Originality Statement

All formal models, terminology, and evaluative frameworks presented in this book are my own original contributions, developed through independent reasoning and research. These include, but are not limited to: the event-valued evaluator, Margan's Optimization Paradox (MOP), Minimal-Trigger Optimality, and the Confirmation operator (S^*), together with the broader Margan Optimization Framework and its guardrails. Associated terminology introduced here — such as scope locks, honesty locks, event ledgers, remediation dockets, and closeout certificates — was likewise formulated during the course of this work. These models and terms are not drawn from existing literature; they have been worked out in the process of my writing and are presented here as original contributions to philosophical debate, applied ethics, and governance practice.

Scripture & Primary Sources

Scripture quotations, where used, are identified by translation in notes or the bibliography. Permissions for specific translations (if any) appear in the bibliography's *Primary Sources* section. Non-scriptural primary standards and statutes are cited to their official sources.

Production & Typesetting

Set in LaTeX using `newtxtext/newtxmath` with microtypographic refinement. Page size: 6 in × 9 in. Figures are embedded as vector PDF where possible. Accessibility features (PDF outlines, tagged structure, alt text) have been prepared to the extent supported by the toolchain.

Identifiers

ISBN (print): [to be assigned] **ISBN (ebook):** [to be assigned]
DOI: 10.5281/zenodo.17079986

Library Cataloguing

A CIP (Cataloguing-in-Publication) record for this book will be available from the national library upon assignment of ISBN.

Contact

Rights, permissions, and desk/exam copies: rights@[your-publisher].com
Press & speaking: press@[your-publisher].com
Errata and updates: `https://[your-site]/TRO/errata`

Cover design: [Name]. *Interior design:* [Name]. *Printed by:* [Printer].

How to Cite This Book

Sergiu Margan (2025). *The Redemption Optimization: Why Evil Exists in a Divine Blueprint for Ultimate Good*. [Publisher], [City]. DOI: 10.5281/zenodo.17079986.

Prefaces

Why two prefaces?

This book speaks to two audiences at once. The first preface is a personal account in plain language. The second is a compact orientation for technical reviewers and replicators. Read one or both; they dovetail with *How to Read This Book*.

Author's Preface (First Person)

I began with a stubborn question: *If God is love, and human freedom is real, how could goodness be secured for eternity without coercion or the possibility of fresh rejection?* I wrote the initial paper to test a hunch: that the answer requires counting *events* in history, not merely capacities or intentions.

The core insight crystalised in what I call **Margan's Optimization Paradox (MOP)**: if we truly value *realized* mercy and justice, then a world with *no* counted rejection minting them will credit *none* of those goods. Under an *event-valued* evaluator, redemption cannot be only a potential—it must interface with an actual wound, once,

and then be *closed* non-coercively. From there emerged a surprising simplicity:

- **Count events.** Credit only what happens in history: delivered protections (ΔL), exercised freedoms (ΔF), enacted mercy and justice (M, J^{v}). Do not mint credit for plans or near-misses.

- **Minimal trigger.** If redemption-goods matter, one gated rejection per kind is sufficient; repeats add cost but no new kind-level value.

- **Confirmation (S^{*}).** Prove the hazardous route is *structurally* closed by cohort replay → stress drills → a monitored live window—while real freedom remains live.

This book is my attempt to bring that logic into the open: faithful to Scripture's moral grammar, precise enough for philosophical scrutiny, and concrete enough for engineers, clinicians, and policymakers who must make systems safer *this week*. It is not a license to tolerate harm. It is a blueprint for *ending* specific harm routes—honestly, with artefacts—while protecting human agency.

I wrote in the first person because the work began as a personal wrestle. I keep it conditional and auditable because truth does not need special pleading; it needs clarity, evidence, and fair tests. If you disagree, I hope you will use the appendices to build a counterexample cleanly. Either the route closes, or it doesn't; either an event minted mercy, or it did not. That is the spirit of the whole project.

Sergiu Margan
August 2025

Technical Preface (For Reviewers & Replicators)

Evaluator (event-valued). We adopt a single objective,

$$\mathcal{J} = \sum_{t \geq 0} \left(\alpha \, \Delta L_t + \gamma \, \Delta F_t + \mu \, M_t + \nu \, J_t^{\nu} \right) \; - \; \sum_{t \geq 0} \left(\beta \, H_t + \kappa \, R_t \right),$$

under guardrails A1–A8 (Apps. A, C): event-valued credit only (A1), non–coercion and bounded harm (A2), culpability gate and harm floor (A3), typal/concave kind-credits (A4), positive costs (A5), feasible branches (A6), structural confirmation operator (A7), and scope lock (A8). A capacity-valued comparator $\mathcal{J}^{\dagger}$ is defined to make divergences explicit.

Main results (conditional). (i) **Two-Branch Fork / MOP:** if $\mu + \nu > 0$ and $\sum_t R_t = 0$, then $\sum_t (M_t + J_t^{\nu}) = 0$; the evil-free branch is coherent only when redemption-goods are not valued in the event sense. (ii) **Minimal-trigger optimality:** with typal/concave credits and positive per-event costs, at most one gated rejection per class before closure maximises $\mathcal{J}$; repetition is strictly dominated. (iii) **Confirmation:** S^* is a structural property (hazard off, freedom live), certified by three stages.

Reproducibility. Ch. 41 and Apps. G–F specify schemas, metrics, ablations, falsifiers, and exact commands to regenerate figures/tables.

Seeds, thresholds, and weights are versioned; checksum manifests are provided. A tagged release (`v1.0-book`) corresponds to this edition; a DOI will be posted in the repository README.

Claims discipline. Results are stated *only* under declared assumptions. Where we switch evaluators (to $\mathcal{J}^{\dagger}$), this is flagged. Each theorem lists exactly which A#'s it uses; reviewers can audit or contest at that interface.

Ethics. Examples are synthetic, anonymised, or public. Event-valued accounting explicitly forbids counting suppression or coerced "closure" as success.

Acknowledgments (Brief)

I am grateful to readers and critics who pressed objections until the definitions were tight, and to practitioners across domains who insist that accountability be proven by events, not promises. Any remaining errors are mine.

How to Cite

Sergiu Margan, *The Redemption Optimization: Why Evil Exists in a Divine Blueprint for Ultimate Good*, v1.0-book (year of this edition). If citing a formal result, include the theorem/appendix label (e.g., "Thm. A.2, App. A").

One-line orientation

Count events, fix routes, prove closure, keep freedom live. The rest of the book makes that sentence precise, testable, and usable.

Note on Terms & Usage

Notation

This page standardises names, symbols, and editorial choices used throughout the book. For full symbol definitions see App. B; for assumptions and proofs see App. A and App. C.

Names you will see

- **Margan's Optimization Paradox** (**MOP**): the *Two-Branch Fork* result—if redemption-goods are valued ($\mu + \nu > 0$), an evil-free history ($\sum_t R_t = 0$) earns *zero* event-valued redemption credit; otherwise at least one realised refusal is necessary to ground mercy/justice. (Formal statement: Thm. A.1.)

- **Margan's Optimization Framework**: the *event-valued* evaluator $\mathcal{J}$ with its guardrails (A1–A8), minimal-trigger doctrine, and confirmation operator S^*. We refer to it simply as the **Framework**.

- **Trademark note.** For readability we omit the ™/® symbols in running text. If you later register a mark, it may appear on the title page or copyright page; the mathematics and definitions are unchanged.

Moral terms mapped to formal objects

- **Love / Protection** $\rightarrow \Delta L_t$ (delivered, *event-valued*).
- **Freedom / Rights exercised** $\rightarrow \Delta F_t$ (delivered, event-valued).
- **Sin / Evil = Rejection** $\rightarrow R_t = 1$ iff (breach $B_t = 1$ or harm $H_t \geq h_{\min}$) *and* culpability $C_t = 1$ (A3). Accidents ($C_t = 0$) may add harm for learning but do not mint R_t.
- **Mercy (restoration)** $\rightarrow M_t \in \{0, 1\}$ when enacted to actual victims (A1).
- **Justice (accountability)** $\rightarrow J_t^v \in \{0, 1\}$ when enacted (A1).
- **Confirmation** (S^*) $\rightarrow$ structural hazard removal: cohort replay $\rightarrow$ stress drills $\rightarrow$ monitored live; on pass, set $S^* = 1$ and hazard HZ = 0 for that class (A7–A8).

Evaluator and guardrails (plain language)

- **Evaluator.** We score *events*:

$$\mathcal{J} = \sum_{t \geq 0} \big(\alpha\, \Delta L_t + \gamma\, \Delta F_t + \mu\, M_t + \nu\, J_t^v\big) - \sum_{t \geq 0} \big(\beta\, H_t + \kappa\, R_t\big),$$

under (i) **non–coercion** (live alternate possibilities), (ii) **bounded harm** $E_{\text{tot}} \leq H_{\max}$, and (iii) **honesty locks** (no suppression).

- **Typal/concave credits.** Kind-level redemption credit saturates at the *first* counted rejection within a class; repeats earn no extra kind-credit (A4).

- **Minimal-trigger doctrine.** With positive costs ($\beta, \kappa > 0$) and typal/concave credits, the best policy is: one counted wound per class → remediate → confirm → lock scope; repetition is dominated (Thm. A.2).

Domain language

- **Class** c = a specific hazardous *causal route* (e.g., `TAB-ERR` in elections, `ALLERGEN` in food safety). After closure, the class definition is *scope-locked*; a genuinely different causal vector opens a new class c' (A8).

- **Docket** = the remediation pack opened at the first validated $R_t^{(c)} = 1$ (root cause, fixes, tests, owner, deadlines, artefacts).

- **Artefact** = reproducible evidence (logs, configs, model cards, SOPs, drill reports, checksums) that a fix happened in history.

Editorial conventions

- **Spelling.** We use *Optimization* (American technical spelling) consistently, including in the title; other prose follows standard international academic English.
- **Scripture.** When Scripture is cited, translations are identified on first use or in the note; references appear as *Book chapter:verse* (e.g., *Romans 5:8*). Primary sources are listed in the bibliography.
- **Typesetting cues.** Binary switches appear in small caps (e.g., HZ, S^*). Scalar variables are italic; sets/operators are calligraphic (e.g., $C, \mathcal{T}, \mathcal{M}$). Near-miss is hyphenated; event-valued/capacity-valued are hyphenated.
- **Ethics of examples.** All case examples are anonymised, public, or synthetic; when partner material informs a template, only aggregate/redacted artefacts are reproduced.

What counts and what does not

- **Counts (event-valued):** delivered protections and freedoms; enacted mercy/justice; adjudicated rejections; structural confirmations with artefacts.
- **Does not count:** intentions, capacities, policies on paper, near-misses (unless eventised as drills; they support S^* but mint no M, J^v).

Quick reminder for readers. When you see M or J^{ν} credited, ask: what event minted it? When you see $S^{*} = 1$, ask: where are the cohort-replay, stress-drill, and monitored-live artefacts? Those questions enforce the event-valued stance throughout this work.

Terms & Attribution

Notation

Purpose. This page clarifies which terms are coined or specialised in this book, what we mean by them, and where we acknowledge prior, widely used terms (laws, standards, acronyms). It also states our practice on trademarks, scripture, and reuse of the book's operational templates.

Coined or specialised terms in this volume (by the author)

Name	Meaning / Use in this book

Margan's Optimization Paradox (MOP)	The Two-Branch Fork: under an *event-valued* evaluator, if redemption-goods are valued ($\mu + \nu > 0$), then an evil-free history ($\sum R_t = 0$) earns zero redemption credit; otherwise at least one realised rejection is required to ground mercy/justice.
Margan's Optimization Framework	The event-valued evaluator $\mathcal{J}$ with guardrails (A1–A8), minimal-trigger doctrine, and structural confirmation (S^*).
Event-valued / Capacity-valued (comparators)	"Event-valued" credits only what happened in history; "capacity-valued" may credit potentials. Used to cleanly separate two evaluation stances.
Minimal-trigger optimality	With typal/concave kind-credits and positive per-event costs, one counted rejection per class before closure strictly dominates repetition.
Confirmation (S^*) as structural hazard removal	A three-stage proof (cohort replay → stress drills → monitored live) that the hazardous route is closed (HZ = 0) while alternate possibilities remain live.
Scope lock	Freezing a class definition after S^*; recurrences via the same causal vector violate closure; genuinely new vectors open a new class c'.

Honesty locks	Anti-suppression practices (random checks, whistleblowing, pre-registered criteria) that protect evidence integrity.
Event ledger	Minimal schema that records $R, B, C, H, \Delta L, \Delta F, M, J^v, \mathrm{HZ}, S^*$ for audit.
Remediation docket / Closeout certificate	Operational artefacts: the docket opened at first $R^{(c)} = 1$ and the one-page closure attestation upon $S^*(c) = 1$.

Acknowledged prior terms (laws, standards, acronyms)

Term	Field	Notes / Attribution
Asimov's "Three Laws of Robotics"	Robotics ethics (literary)	Coined by Isaac Asimov; referenced here only as a cultural touchstone.
Goodhart's Law	Economics/measurement	Attributed to Charles Goodhart; "when a measure becomes a target, it ceases to be a good measure." Used for evaluation caution.
Risk-Limiting Audit (RLA)	Elections	Term of art in election auditing; used here in its standard sense.
ICS / EOC	Emergency management	Incident Command System / Emergency Operations Center; standard organisational terms.

HACCP / FSMS	Food safety	Hazard Analysis & Critical Control Points / Food Safety Management System; standard frameworks.
SPC / CUSUM / EMA	Quality engineering	Statistical Process Control; cumulative sum; exponential moving average; standard control tools.
KYC / CDD / AML / CFT	Finance	Know Your Customer; Customer Due Diligence; Anti-Money Laundering; Counter-Terrorist Financing; regulatory terms.
SCADA / Black-start	Energy/utilities	Supervisory Control and Data Acquisition; grid restart capability; industry terms.
RLA, L&A	Elections	Risk-Limiting Audit; Logic & Accuracy testing; standard practices.

Trademarks and proper names

All product, standard, and company names mentioned are the property of their respective holders and are used in a descriptive, nominative sense. We omit ™/® symbols in running text for readability (see Publisher's Note). If you later register a mark related to this work, it may appear on the title/copyright pages without altering the mathematics.

Scripture and primary sources

Where Scripture is quoted, the translation is identified in notes or in the bibliography's "Primary Sources" subsection. Citations to laws, standards, or public guidance are to their official sources as listed in the bibliography.

Reuse of templates and figures

Operational templates (scope locks, dockets, confirmation plans, posters) and code-generated figures/tables are reproducible from the companion repository. Reuse follows the licences stated in that repository's `LICENSE` file. Book prose, page layouts, and non-code artwork remain © the author/publisher (see Publisher's Note).

How to request a correction

If you believe a term has been misattributed or should be credited differently, please email the address on the Publisher's Note with a short justification; corrections will be posted on the public errata page in subsequent printings.

Editorial practice in one line: we coin terms only where needed, we acknowledge prior art explicitly, and we keep trademarks and scripture references clear and fair.

Part I

The Problem Parameters

Chapter 1

The Terrain of the Problem

Notation

Core symbols (used informally in this chapter; formalized later).

$R_t \in \{0,1\}$ (rejection event at time t); $B_t \in \{0,1\}$ (rights/consent breach); $C_t \in \{0,1\}$ (culpability gate); $H_t \geq 0$ (realized harm); $\mathrm{HZ}_t \in \{0,1\}$ (hazard on/off); $S_t^* \in \{0,1\}$ (confirmation);

$E_{\mathrm{tot}} = \sum_t H_t$ (cumulative realized harm); $\mathcal{J}$ (the event-valued evaluator).

A first-person starting point

I began with a stubborn question: *If love and freedom are genuine, can history reach a future where rejection never happens again, without erasing freedom?* The familiar responses—"evil is the price

of freedom," or "this is the best possible world"—felt too loose for a claim about reality. I wanted something I could *count*: events, not intentions; repairs, not rhetoric. That push toward countable, verifiable history is what led me to an *event-valued* approach.

Why events (not capacities) count here

Many frameworks reward *capacities* (what could have happened) instead of *events* (what actually did). That distinction decides everything in this book:

- **Capacity-valued** talk may credit "mercy" or "justice" because they *could* occur, even if nothing happened.
- **Event-valued** evaluation credits mercy or justice *only when enacted on the realized path.*

Choosing event-valued evaluation prevents us from smuggling in unrealized goods and forces intellectual honesty about what history has, in fact, borne.

The problem, framed

We care about three families of goods measured on history:

- **Love-goods** (L): reconciliations, safeguarding actions, sustained care delivered.
- **Freedom-goods** (F): live alternate possibilities (AP) exercised without coercion.

- **Redemption-goods**: realized *mercy* (M) and realized *justice* (J^v) when a real wrong has occurred and is rightly addressed.

We also track realized harms H_t and insist on guardrails: non–coercion, bounded harm with redress, and a structural *confirmation* phase (S^*) in which the hazardous route is truly closed while freedom remains live.

Rejection (event-level)

Let $R_t \in \{0,1\}$, $B_t \in \{0,1\}$, $H_t \geq h_{\min} \geq 0$, and $C_t \in \{0,1\}$. We say

$$R_t = 1 \text{ iff } \big(B_t = 1 \text{ or } H_t \geq h_{\min}\big) \text{ and } C_t = 1.$$

Here $C_t = 1$ is the *culpability gate* (intentional / reckless / grossly negligent). Accidents keep $C_t = 0$ (not counted as rejection), though their harm still informs safety learning.

Two-branch fork (the intuitive picture)

- **Evil-free branch.** Everyone freely chooses YES from the start; realized mercy/justice remain 0 because there is no wrong to address. Coherent under a capacity-valued lens or an event-valued lens that sets the redemption weights $\mu = \nu = 0$.

- **Redemptive branch (minimal trigger).** A *single* realized refusal occurs ($\sum_t R_t = 1$), followed by consent, mercy and justice enacted, a *hazard reset* (HZ = 0), and structural

confirmation $S^* = 1$ so that future rejection is volitionally absent without coercion.

The heart of this book is to make that fork precise, then show why—*if you value realized redemption-goods at all* ($\mu, \nu > 0$)—the minimal-trigger redemptive arc is Pareto-optimal and repetition of the same wound is dominated once costs are counted.

A preview of the evaluator

We will evaluate on the realized path (not on mere potential):

$$\mathcal{J} = \sum_{t \geq 0} \Big(\alpha \, \Delta L_t + \gamma \, \Delta F_t + \mu \, M_t + \nu \, J_t^{v} \Big) \; - \; \sum_{t \geq 0} \big(\beta \, H_t + \kappa \, R_t \big), \tag{1.1}$$

with simple state variables and switches:

$$\mathrm{HZ}_t \in \{0, 1\}, \qquad S_t^* \in \{0, 1\}, \qquad R_t, M_t, J_t^{v} \in \{0, 1\}. \tag{1.2}$$

Intuitively: we *add* realized goods ($\Delta L_t, \Delta F_t, M_t, J_t^{v}$) and *subtract* realized costs (H_t and a penalty κ for each rejection). We later prove that, under typal/concave credits for M, J^{v} and $\kappa > 0$, repeats of the *same* wound add no new kind of value and are strictly dominated by a one-off, repaired refusal.

Guardrails (informal statement)

Throughout, we impose four guardrails:

1. **Non-coercion.** Freedom means live AP in the present; confirmation cannot be mind control.

2. **Bounded harm with redress.** Harms are repaired to those actually harmed; cumulative harm cannot diverge.

3. **Honesty locks.** No counting goods that did not happen; no re-labeling repeats as "new kinds".

4. **Structural confirmation.** After remediation, the hazardous route is *closed* (HZ = 0) while live freedom remains.

What follows (chapter map)

- **Part I** situates the work: prior theodicies, why they help, and where they under-specify objectives/constraints.

- **Part II** gives precise definitions, the evaluator, and the fork as lemmas; it proves the *Margan Optimization Paradox* (MOP) in our vocabulary: eternal love, live AP at every moment, and zero rejection at every moment cannot all hold together.

- **Part III** proves *minimal-trigger optimality* under event-valued redemption and shows why repeats are dominated with costs.

- **Part IV** operationalizes the ledger in AI safety and governance with audits, dockets, tests, and escalation ladders.

Contributions (informal)

1. **Event-valued evaluator.** A ledger that attaches credit/debit to what actually happens.

2. **Conditional necessity.** If you value realized mercy/justice ($\mu, \nu > 0$), at least one realized refusal is necessary to realize them.

3. **Minimal-trigger dominance.** With typal/concave credits and $\kappa > 0$, repeated wounds do not increase $\mathcal{J}$ once costs are counted; the one-off, repaired route is Pareto-optimal.

4. **Confirmation as structure.** S^* is not coercion but a stable closure of the hazardous route with live freedom.

5. **Governance translation.** A practical playbook: incident classes, remediation dockets, confirmation tests, and honest metrics.

A small formal teaser

Theorem 1.1 (Two-Branch Preview)**.** *Under the event-valued evaluator* (1.1) *with weights* $\mu, \nu \geq 0$*, the evil-free branch yields* $M_t = J_t^{\nu} = 0$ *for all* t*. If* $\mu, \nu > 0$*, any policy that realizes* (M, J^{ν}) *on history must realize at least one* $R_t = 1$*.*

Idea of proof. On the evil-free branch there is no wrong to remedy, so (M, J^{ν}) remain zero by definition. If (M, J^{ν}) are credited only when enacted on realized history, some R_t must occur to ground

them; otherwise the credit would violate the honesty lock. Full proofs appear in Part II.

Reading tip. If you prefer the narrative first, skip to Part IV and return to the formal results later; the book is written to support both paths.

Chapter 2

Theodicies and Their Shortfalls

Notation

Scope of this chapter. We map classic responses to the problem of evil and identify (i) what each explains well, and (ii) what remains underspecified for a model that evaluates *events* on history (not mere capacities). Formalism begins in Part II; here we keep symbols light and intuitive.

Why review the tradition at all?

Our evaluator does not start from scratch. It keeps the insights of classic theodicies but demands precision about *what gets counted, when, and why*. In particular, we separate:

- **Capacity-valued claims** (“could have been merciful”) from
- **Event-valued facts** (“mercy actually enacted here and now”).

This distinction will quietly decide many disagreements.

Five influential lines

1) Free-Will Defense. Evil is possible because creatures have genuine alternate possibilities (AP). This preserves agency and avoids global determinism.

- *Gets right:* Agency matters; permission of refusal is intelligible.
- *Shortfall:* No built-in convergence to a future with live freedom *and* no rejection; no limit on cumulative harm; no accounting for *realized* mercy/justice.

2) Soul-Making. Hardship can develop virtues (courage, compassion) unreachable in a frictionless world.

- *Gets right:* Developmental goods are real.
- *Shortfall:* Without guardrails, can drift into justifying excessive harms; lacks a *hazard reset* that prevents repetition; often pays out virtue credits even when no repair happened in history.

3) Best Possible World / Greater-Good. God permits certain evils because they are outweighed by greater goods.

- *Gets right:* Trade-offs exist; optimization language is natural.
- *Shortfall:* Objective, weights, and constraints are rarely formalized; can justify anything without honesty locks; often counts modal goods that never occurred.

4) Skeptical Theism. Our cognitive limits prevent us from seeing God's justifying reasons.

- *Gets right:* Epistemic humility.
- *Shortfall:* Non-constructive; does not yield a *testable* evaluator or policy; cannot by itself secure an eternal order with agency and bounded harm.

5) Narrative/Redemption Theodicies. History contains a redemption arc in which wrong is addressed and transformed.

- *Gets right:* Mercy and justice are goods that *arrive in history*.
- *Shortfall:* Often evaluated narratively, not with an explicit ledger that penalizes real harm and prevents repetition.

A side-by-side at a glance

Approach	What it gets right	Where it falls short vs. our target
Free-Will Defense	Preserves agency; explains permission of wrong	No built-in convergence to a no-rejection end-state with live freedom; no bound on cumulative harm; no event-valued ledger for mercy/justice
Soul-Making	Values developmental goods	Lacks hazard reset; risks excessive harm; often counts virtue without repair on the realized path
Best Possible World / Greater-Good	Recognizes trade-offs and optimization language	Objective/weights/constraints underspecified; risks licensing almost anything without honesty locks; modal goods over-counted
Skeptical Theism	Restores epistemic humility	Non-constructive; offers no testable evaluator; cannot alone yield an eternal order with agency and bounded harm
Narrative/Redemption	Sees mercy and justice as historical goods	Rarely penalizes realized costs explicitly; lacks confirmation tests to prevent recurrence

What we keep—and what we add

We keep the insights (agency matters; development is real; trade-offs exist; humility is wise; redemption is central). We *add*:

1. an **event-valued ledger** that credits mercy/justice only when enacted;

2. **guardrails**: non–coercion, bounded harm with redress, honesty locks, and structural confirmation (S^*);

3. a **minimal-trigger** doctrine: if wrong occurs, one refusal is enough for all redemption-goods of that type; repeats add costs but no new kind of value under typal/concave credits.

Event-valued vs capacity-valued (worked definition)

Event-valued vs. capacity-valued evaluation

Capacity-valued evaluation assigns credit for goods that *could* occur given abilities or opportunities. *Event-valued* evaluation assigns credit only to goods *that did occur* on the realized path. In this book, redemption-goods (mercy, justice) are event-valued: they require an actual wrong addressed in history. This is the pivotal assumption that makes the fork non-trivial.

Where the classic views leave gaps (for us to fill)

- **Counting rule.** Which goods are credited, and when? (We make this explicit.)
- **Costs and caps.** How are realized harms penalized and bounded? (We set E_{tot} bounds and redress rules.)
- **Convergence.** What ensures a future with live freedom and no renewed rejection? (We formalize S^* as structural hazard removal.)
- **No repetition bonus.** Why don't repeats of the same wound count as new goods? (Typal/concave credits + explicit penalties.)

How this sets up Part II

Part II will introduce the evaluator $\mathcal{J}$ rigorously, define the state variables ($R_t, B_t, C_t, H_t, \text{HZ}_t, S_t^*$), prove the *Margan Optimization Paradox* (MOP), and present the *Two-Branch Lemma*. With that, the intuitive critiques above become formal statements with testable implications.

Chapter 3

Defining the Goods

Notation

State and evaluand (informal preview).
L_t (love-goods increment), F_t (freedom-goods increment), M_t (realized mercy), J_t^y (realized justice), H_t (realized harm), R_t (rejection event), $\mathrm{HZ}_t \in \{0,1\}$ (hazard on/off), $S_t^* \in \{0,1\}$ (confirmation switch), $E_{\mathrm{tot}} = \sum_{t\geq 0} H_t$, and evaluator $\mathcal{J}$ on the realized path.

What we are measuring (event-valued)

We evaluate goods and costs on what *actually occurs* in history. This chapter fixes the meanings of the goods that enter the objective and the guardrails that constrain policies.

Love-goods L (event increments)

$L_t \geq 0$ records concrete goods of love realized at time t (e.g., verified reconciliation, safeguarding action completed, sustained care delivered). We write ΔL_t for the increment credited at t to avoid double-counting long projects.

Freedom-goods F (event increments)

$F_t \geq 0$ records realized exercises of freedom with live alternate possibilities (AP) and without coercion (e.g., rights actually exercised, due process met). We credit *events*, not mere capacities.

Redemption-goods (event-valued)

$M_t \in \{0, 1\}$ marks realized *mercy* at t (pardon, restoration enacted to an actually harmed party); $J_t^y \in \{0, 1\}$ marks realized *justice* (rectification, restitution, or accountability enacted). Both require a real wrong on the path and cannot be credited modally.

Rejection and harm

Let $R_t \in \{0, 1\}$, $B_t \in \{0, 1\}$ (rights/consent breach), $C_t \in \{0, 1\}$ (culpability gate), and $H_t \geq 0$ (realized harm). We set

$$R_t = 1 \text{ iff } \big(B_t = 1 \text{ or } H_t \geq h_{\min}\big) \text{ and } C_t = 1.$$

Accidents keep $C_t = 0$; their H_t still informs safety learning but is not counted as rejection.

Guardrails (assumptions used later)

title=A1 — Non-coercion

Live AP in the present; no brainwashing, no lobotomy solutions. Confirmation (later S^*) must not negate agency. Live AP in the present; no brainwashing, no lobotomy solutions. Confirmation (later S^*) must not negate agency.

title=A2 — Bounded harm with redress

There exists $H_{\max} < \infty$ such that $E_{\text{tot}} \leq H_{\max}$ along admissible policies, and redress is directed to the actually harmed party.

title=A3 — Honesty locks (event-valued credits)

No credit for goods that did not occur on the realized path; no counting repeats of the same wound as a new kind of good.

title=A4 — Structural confirmation

There is a policy phase (denoted by $S^* = 1$) after which the hazardous route is *closed* (HZ = 0) while freedom remains live.

Objective (illustrative form; split over lines to fit)

We adopt an event-valued evaluator with additive goods and explicit costs:

$$\mathcal{J} = \sum_{t \geq 0} \Big(\alpha\, \Delta L_t + \gamma\, \Delta F_t + \mu\, M_t + \nu\, J_t^{v} \Big) \; - \; \sum_{t \geq 0} \Big(\beta\, H_t + \kappa\, R_t \Big), \tag{3.1}$$

$$\text{s.t.} \quad \mathrm{HZ}_t \in \{0,1\},\; S_t^{*} \in \{0,1\},\; E_{\text{tot}} = \sum_{t \geq 0} H_t \leq H_{\max},$$

$$\Big(\textstyle\sum_t R_t \geq 1 \;\&\; \text{remediation passes} \Big) \;\Rightarrow\; (S^{*} = 1,\; \mathrm{HZ} = 0 \text{ thereafter}).$$

Weights $(\alpha, \gamma, \mu, \nu, \beta, \kappa)$ may be normalized; what matters is that $\kappa > 0$ (rejections carry intrinsic cost) and M_t, J_t^{v} are *event-valued*.

Typal / concave credits (no repetition bonus)

To prevent a perverse incentive to *farm* wounds, we use concave "credit schedules" for redemption-goods:

$\mathrm{Credit}_{\text{type } c}(n)$ is increasing and concave in n, $\quad \lim_{n \to \infty} \Delta\mathrm{Credit}(n) = 0.$

In the typal limit, the *first* instance of a wound-type c earns the only kind-level credit; repeats add cost but no new kind of value.

Two canonical micro-timelines (state-level)

$\mathbf{H}_0$ — Evil-free branch (coherent if $\mu = \nu = 0$). All agents freely choose YES. No wrong occurs; hence $M_t = J_t^v = 0$ for all t. Hazard is removed immediately (HZ = 0), and $S^* = 1$ by construction.

$\mathbf{H}_1$ — Minimal-trigger redemptive branch (if $\mu, \nu > 0$). One realized refusal occurs ($\sum_t R_t = 1$), followed by consent and enacted mercy/justice ($M, J^v = 1$ at some times). A remediation passes, hazard class is closed (HZ = 0), and structural confirmation sets $S^* = 1$ with live freedom.

Measurement notes (how to count)

- ΔL_t is credited when reconciliation/safeguarding is *verified* (documented completion, not mere intention).

- ΔF_t is credited when rights are *actually exercised* (e.g., appeals upheld, access granted), not merely available.

- M_t, J_t^v are credited on adjudicated events and delivered remedies; narrative claims are insufficient without event documentation.

- H_t is recorded from incident logs, clinical harm scales, or judged loss; thresholds $h_{\min}$ must be published.

Why these choices matter

With A1–A4 and (3.1), the evaluator forbids two failures: (i) *over-crediting* goods that never happened, and (ii) *under-penalizing* repeats that re-open hazards. This sets up the formal fork in Part II and the *Margan Optimization Paradox* in Chapter 7.

Chapter 4

Interlude I — Literature Review

Notation

What this chapter does. It orients the reader in five literatures that touch our evaluator: (i) classic theodicies, (ii) modern analytic philosophy of religion, (iii) ethics/decision theory/control, (iv) optimization and systems safety (incl. Goodhart effects), and (v) AI safety/evaluation. We highlight what each gets right and where an *event-valued* ledger (our approach) fills explicit gaps: counting only goods enacted in history, penalizing realized costs, and enforcing guardrails (non–coercion, bounded harm with redress, structural confirmation S^*).

4.1 Classical and early-modern backdrop (very brief)

From early statements of the logical and evidential problems of evil to responses that appeal to divine goodness and providence, the tradition offers durable ideas we will keep:

- **Agency matters.** Human (and angelic) freedom is not an illusion; refusal must be a live possibility (alternate possibilities, AP).
- **Development matters.** Virtues often grow through friction and adversity.
- **Trade-offs exist.** Even an all-wise governance of history may involve balancing goods under constraints.
- **Humility matters.** Finite agents may not see all justifying reasons.

What is typically *under-specified* are the objective, weights, constraints, and *accounting rules* by which goods and costs are actually tallied on history.

4.2 Twentieth-century moves in philosophy of religion

Several influential lines sharpen the debate:

- **Free-will defenses** safeguard AP to block logical contradictions.

- **Soul-making** emphasizes developmental goods (character, courage, compassion).
- **Best-possible-world / greater-good** frames justify permissions by outweighing goods.
- **Skeptical theism** urges epistemic humility regarding God's reasons.
- **Narrative/redemption theodicies** locate meaning in a history that is repaired and transfigured.

We adopt insights from each but require a ledger that credits *only realized goods* and explicitly penalizes realized costs, with guardrails.

4.3 Decision theory, control, and optimization (bridges)

Outside theology, adjacent fields teach us to be explicit:

- **Objectives and constraints.** If you do not write them down, you will smuggle them in implicitly.
- **Safety vs. performance.** Control and assurance separate nominal goals from safety envelopes and hazard states.
- **Goodhart's law.** When a measure becomes a target, it can be gamed; honesty locks and confirmation tests are needed.
- **Structural fixes.** Durable closure of a failure mode is different from ad hoc patching; this motivates our S^* notion.

4.4 AI ethics and evaluation (where we will apply the ledger)

Modern AI practice already uses ingredients we recast in event-valued form:

- **Audits and incident response.** Red-teaming, postmortems, and mitigations are common, but often lack a principled tally of *realized* goods and costs.

- **Alignment and governance.** RLHF, policies, and model cards speak to *intentions*; we insist on *events*.

- **Confirmation tests.** Sandboxes and reenactments exist, but we require an explicit *hazard reset* evidenced by tests before returning to scale.

4.5 Side-by-side comparison

Line of thought	Core contribution	Gap vs. event-valued target
Free-will defense	Preserves genuine AP and intelligible permission	No built-in convergence to no-rejection with live freedom; weak treatment of realized costs and redemption accounting
Soul-making	Values developmental goods	Risks over-permitting harm; lacks hazard reset; often credits virtue without event-level repair
Best possible world / greater good	Puts trade-offs on the table	Objective/weights/constraints left implicit; modal goods over-counted; no honesty locks
Skeptical theism	Epistemic humility	Non-constructive for evaluation/policy; cannot itself yield guardrails or confirmation
Narrative/redemption	Sees mercy/justice as historical goods	Rarely uses a ledger that penalizes cost and prevents repetition explicitly
Decision/control/optimization	Precision about goals, safety envelopes	Needs theological semantics for goods like love, mercy, justice; lacks moral culpability gates

4.6 Why an event-valued ledger (and what it adds)

Our evaluator $\mathcal{J}$ credits only goods *enacted on the realized path* (love increments ΔL_t, freedom increments ΔF_t, realized mercy M_t, realized justice J_t^{ν}) and subtracts *realized* costs (harm H_t and a rejection penalty κR_t). Guardrails include non–coercion, bounded harm with redress, honesty locks, and structural confirmation (S^*). This yields:

1. **Counting discipline.** No credit for unrealized "could-have" goods.

2. **Cost visibility.** No hiding realized harm inside grand narratives.

3. **No repetition bonus.** Typal/concave credits prevent "farming" wounds.

4. **Closure requirement.** Return to scale only after a passed confirmation test (hazard class closed, HZ = 0).

4.7 How MOP and MOF sit in this landscape

MOP (Margan's Optimization Paradox). Informally: eternal love, live AP at every moment, and zero rejection at every moment cannot all hold together. In event-valued terms, either redemption-goods stay at 0 forever (evil-free branch), or at least one realized refusal grounds them before structural confirmation.

MOF (Margan's Optimization Framework). The event-valued evaluator with guardrails and the minimal-trigger, confirmation-based arc. If $\mu, \nu > 0$ (you value realized mercy/justice), one realized refusal is *necessary and sufficient* for the kind-level credits; repeats add cost but no new kind of value once typal/concave credits and $\kappa > 0$ are in place.

4.8 What we will prove next (reader's map)

- In Part II we formalize the state, the evaluator, and the two-branch lemma, then prove MOP precisely.
- In Part III we prove minimal-trigger Pareto-optimality and show why repetition is dominated when costs are counted.
- In Part IV we translate the ledger into AI governance and sector policy with playbooks and confirmation tests.

4.9 Minimal bibliography guide (for your References page)

This interlude is intentionally citation-light for clean compilation. Add your preferred references (e.g., classic treatments of the problem of evil, free-will defenses, soul-making accounts, decision/control texts, Goodhart's law discussions, and AI evaluation papers) to your `references.tex` as you finalize.

Chapter 5

Interlude II — Methods and Modeling Choices

Notation

Purpose. This chapter fixes the modeling choices that make later claims testable: (i) *event-valued* rather than capacity-valued credits, (ii) explicit guardrails, (iii) measurement rules for goods and harms, (iv) culpability gates, (v) handling near-misses and drills, and (vi) reproducibility protocols.

Event-valued evaluation (counting rule)

Event-valued vs. capacity-valued

Capacity-valued evaluation assigns credit for goods that *could* occur given abilities or opportunities. *Event-valued* evaluation assigns credit only to goods that *did occur* on the realized path. In this book, redemption-goods (realized mercy M_t and realized justice J_t^v) are *event-valued*.

We thus forbid crediting unrealized "could-have" goods, and we *subtract* realized costs. The evaluator is used descriptively in Parts II–III and prescriptively in Part IV.

State, events, and switches (informal preview)

Core variables

At time t:

$R_t \in \{0,1\}, \quad B_t \in \{0,1\}, \quad C_t \in \{0,1\}, \quad H_t \geq 0, \quad M_t, J_t^v \in \{0,1\}, \quad \Delta L_t, \Delta F_t$

$\mathrm{HZ}_t \in \{0,1\}$ (hazard on/off), $\quad S_t^* \in \{0,1\}$ (confirmation switch), $\quad E_{\mathrm{tot}} = \sum_{t \geq 0}$

Guardrails (assumptions used by later theorems)

We phrase guardrails as labeled assumptions we can cite in proofs.

A1 — Non-coercion (live AP)

Freedom means live alternate possibilities in the present. Confirmation (S^*) is not mind control; it is a structural change that removes a route without negating agency.

A2 — Bounded harm with redress

There exists $H_{\max} < \infty$ such that along admissible policies $E_{\text{tot}} \le H_{\max}$. Redress is directed to the actually harmed party; aggregate offsets do not erase personal harms.

A3 — Honesty locks (event-valued credits)

No credit for goods that did not occur on the realized path; no new *kind-level* credit for repeating the same wound. Typal/concave schedules apply to M, J^v.

A4 — Structural confirmation

After remediation passes for an incident class, set $S^* = 1$ and force HZ = 0 for that class thereafter (route closed). Live freedom remains.

Objective (illustrative; line-broken to fit)

We use an event-valued objective with explicit costs:

$$\mathcal{J} = \sum_{t \ge 0} \Big(\alpha\, \Delta L_t + \gamma\, \Delta F_t + \mu\, M_t + \nu\, J_t^v \Big) \; - \; \sum_{t \ge 0} \Big(\beta\, H_t + \kappa\, R_t \Big), \tag{5.1}$$

$$\text{s.t.} \quad \mathrm{HZ}_t \in \{0,1\},\ S_t^* \in \{0,1\},\ E_{\text{tot}} \le H_{\max},$$
$$\Big(\textstyle\sum_t R_t \ge 1 \ \&\ \text{remediation passes}\Big) \Rightarrow \big(S^* = 1,\ \mathrm{HZ} = 0 \text{ thereafter}\big).$$

Only the *signs and roles* matter for our results: $\kappa > 0$ penalizes rejection; M, J^v are event-valued.

Culpability gate (fairness and precision)

Rejection R_t and culpability C_t

Let $h_{\min} \ge 0$ be the harm threshold and $B_t \in \{0,1\}$ a rights/consent breach indicator. Then

$$R_t = 1 \ \text{ iff } \ \big(B_t = 1 \text{ or } H_t \ge h_{\min}\big) \ \text{ and } \ C_t = 1.$$

$C_t = 1$ if (i) adequate knowledge, (ii) adequate freedom, and (iii) at least one fault threshold holds: *intentional*, *reckless*, or *grossly negligent*. If $C_t = 0$ (accident), record H_t for learning but do not set $R_t = 1$.

Near-misses and drills (so learning still counts)

- **Near-miss.** No B_t, $H_t < h_{\min} \Rightarrow$ not a rejection ($R_t = 0$). Still record features for prevention.
- **Eventized drill.** High-fidelity simulation/red-team that stands in for a dangerous scenario; mark a drill flag $D_t = 1$. Drills earn learning credit (toward ΔL_t) but not M_t or J_t^v.

Typal / concave credits (no repetition bonus)

To block perverse incentives to "farm" wounds, redemption credits are concave in the count of *the same wound type*. In the typal limit, the *first* instance earns the kind-level credit; repeats add cost but no new kind of value.

Measurement rules (what evidence qualifies)

- ΔL_t **(love).** Credit when reconciliation/safeguarding is *verified* (documented completion, sustained delivery).

- ΔF_t **(freedom).** Credit when rights are *actually exercised* (access granted; appeal upheld), not merely available.

- M_t **(mercy).** Credit when pardon/restoration is enacted to the actually harmed party with consent logs.

- J_t^v **(justice).** Credit when rectification/accountability is enacted (restitution delivered; sentence served; policy corrected).

- H_t **(harm).** Record from adjudicated incidents, clinical scales, or verified loss; publish $h_{\min}$.

- R_t **(rejection).** Set by the gate: rights breach or harm over threshold *and* culpability $C_t = 1$.

Hazard classes, remediation, confirmation

Incident classes and closure

Partition incidents into classes $c \in C$ that share a causal route. For class c, a remediation docket must specify: *root cause*, *fix*, *tests*, *owner & deadline*, and *evidence to publish*. After a passed confirmation test that reenacts the prior causal route, set $\mathrm{HZ}(c) = 0$ for that class and mark $S^*(c) = 1$.

Minimal reproducibility protocol (what to publish)

- **Data & code.** Provide inputs, scripts, and seeds to reproduce figures/tables.
- **Config.** Publish $h_{\min}$, $H_{\max}$, and κ; document units for H_t.
- **Audit artifacts.** Include sample remediation dockets and confirmation test checklists.
- **Replication note.** Invite an external reproducer to run the scripts and sign a one-paragraph replication note included in an appendix.

Policy-neutral parameterization (defaults to tune)

Parameter	Meaning and default guidance
$h_{\min}$	Harm threshold for counting a rejection; choose per domain (e.g., clinical sentinel thresholds; rights breach per se sets $B_t = 1$).
$H_{\max}$	Cumulative-harm cap for admissible policies; sets escalation triggers.
κ	Intrinsic penalty per $R_t = 1$; ensure $\kappa > 0$ to avoid perverse incentives.
μ, ν	Weights for realized mercy and justice; if either > 0, realized refusals are required to ground credits.
α, γ	Weights for love/freedom increments; tuned to domain priorities with public rationale.

What this enables downstream

With A1–A4, the counting rule, and (5.1) in place, later chapters can: (i) state the two-branch lemma precisely, (ii) prove the *Margan Optimization Paradox* (MOP), and (iii) establish *minimal-trigger* Pareto-optimality with typal/concave credits and $\kappa > 0$. Part IV then operationalizes these with audits, playbooks, and confirmation tests.

Chapter 6

Interlude III — Foundations

Notation

Aim of this interlude. Fix the conceptual pillars we use in the formal parts: (i) event-valued vs. capacity-valued evaluation, (ii) the Two-Branch Fork, (iii) minimal-trigger optimality intuition, (iv) confirmation as *structural* hazard removal (not coercion), (v) assumptions and honest limits (including how our view can be wrong). Symbols previewed (formalized later): $R_t, B_t, C_t \in \{0,1\}$; $H_t \geq 0$; $M_t, J_t^v \in \{0,1\}$; $\Delta L_t, \Delta F_t \geq 0$; $\mathrm{HZ}_t \in \{0,1\}$; $S_t^* \in \{0,1\}$; $E_{\mathrm{tot}} = \sum_t H_t$; evaluator $\mathcal{J}$.

6.1 Event-valued reality vs capacity-valued talk

Counting rule (event-valued vs capacity-valued)

Capacity-valued evaluation assigns credit for goods that *could* occur given abilities or opportunities.
Event-valued evaluation assigns credit only to goods that *did occur* on the realized path.
In this book, redemption-goods (realized mercy M_t and realized justice J_t^v) are **event-valued**.

Why this matters:

- It forbids "smuggling" unrealized mercy/justice into scores.
- It aligns moral claims with verifiable history (audits, records, remedies).
- It makes the core fork substantive: if M, J^v only credit when enacted, an evil-free history keeps them at 0.

Lemma 6.1 (Preview: grounding requirement for M, J^v)**.** *If M_t, J_t^v are event-valued and credited only when enacted on the realized path, then on any history with $R_t = 0$ for all t, we must have $M_t = J_t^v = 0$ for all t.*

Idea. With no wrong on the path there is nothing to pardon or rectify; any positive credit would violate the counting rule ("honesty lock"). Full proofs appear in Part II.

6.2 The Two-Branch Fork (intuitive picture)

Fork summary

Evil-free branch ($\mathbf{H}_0$). Everyone freely chooses YES from the start; no realized wrong $\Rightarrow M_t = J_t^v = 0$. Coherent if you set $\mu = \nu = 0$ (you do not value realized redemption-goods) or work under a capacity-valued lens.

Redemptive branch ($\mathbf{H}_1$). A *single* realized refusal occurs ($\sum_t R_t = 1$), then consent, enacted mercy/justice, hazard reset HZ = 0, and structural confirmation $S^* = 1$ so future rejection becomes volitionally absent without coercion.

A schematic (line-broken to fit the page):

H_1 : $R \Rightarrow$ YES $\Rightarrow (M{+}J^v) \Rightarrow$ remediation passes $\Rightarrow$ HZ = 0 $\Rightarrow S^* = 1$ (freedom l

6.3 Minimal-Trigger Optimality (why one is enough)

Typal/concave credits

Redemption credits are *concave* in the count of the *same wound type*; in the typal limit the first instance earns the kind-level credit and repeats add no new kind of value. Rejections incur an intrinsic penalty $\kappa > 0$ and realized harms H_t accumulate in E_{tot}.

Theorem 6.1 (Preview: minimal-trigger dominance)**.** *Under event-*

valued redemption ($\mu, \nu > 0$), typal/concave credits for M, J^{ν}, and $\kappa > 0$, policies that realize more than one rejection of the same type *are strictly dominated (lower $\mathcal{J}$) by a policy that realizes a* single *rejection of that type followed by consent, remedy,* HZ = 0*, and* $S^{*} = 1$.

Idea. After the first instance, additional instances add costs (κ and H_t) but, by concavity/typal counting, add no new kind-level credit. So repeats can only reduce $\mathcal{J}$. Formal statements and proofs come in Part III.

6.4 Confirmation = structural hazard removal (not coercion)

Structural confirmation

For an incident class c, after remediation passes and reenactment tests succeed, we set $S^{*}(c) = 1$ and force $\mathrm{HZ}(c) = 0$ thereafter (route closed). Freedom remains live: multiple permissible actions exist, but the former hazardous route is structurally unavailable or stably unattractive given the repaired system and formed character.

Practical signals of structural closure:

- Causal route is eliminated (design change, gate, or policy that prevents the prior failure mode).
- Independent confirmation test reenacts the prior scenario with stronger incentives; no recurrence.

- Longitudinal drift checks show stability (no relapse via the same vector).

6.5 Guardrails (assumptions we will cite)

A1 — Non-coercion

Live alternate possibilities (AP) in the present; no brainwashing/lobotomy solutions. S^* cannot negate agency.

A2 — Bounded harm with redress

There exists $H_{\max} < \infty$ with $E_{\text{tot}} \leq H_{\max}$ for admissible policies, and redress is directed to those actually harmed.

A3 — Honesty locks

No credit for goods that did not occur on the realized path; no repetition bonus for the same wound type (concave/typal credits).

A4 — Structural confirmation

Return to scale only after a passed confirmation test; set $S^* = 1$ and HZ = 0 for the closed class.

6.6 Limits, alternatives, and how this could be wrong

Alternative evaluator (capacity-valued). If one adopts a capacity-valued evaluator (call it $\mathcal{J}^{\dagger}$) that credits "could-be mercy/justice" without realized wrong, the evil-free branch can achieve those credits; then our conditional necessity result does *not* apply. The disagreement is about *what counts*.

Domain limits. Our guardrails presuppose: measurable harms, adjudication of rights/consent breaches, and feasible confirmation tests. In opaque domains lacking these, claims should be weakened or withheld.

Falsifiability (what would disconfirm us).

- A documented case where repeating the same wound type *increases* the evaluator despite $\kappa > 0$ and concave credits.
- A coherent, non-coercive S^{*} failure: remediation "passes," yet reenacted tests repeatedly refail via the same causal route.
- Crediting M, J^{v} on an event-valued rule without any realized wrong (violates honesty locks).

6.7 Roadmap to the formal parts

- Part II pins down the state, the evaluator, and proves the **Two-Branch Lemma** and **Margan's Optimization Paradox**

(MOP) in our vocabulary: eternal love, live AP at every moment, and zero rejection at every moment cannot all hold together.

- Part III proves **minimal-trigger optimality** under typal/concave credits and $\kappa > 0$, and characterizes S^* as structural closure with live freedom.
- Part IV operationalizes the ledger for AI/governance with incident classes, remediation dockets, confirmation tests, and escalation ladders.

Part II

The Model (Formal Framework)

Chapter 7

Agents and the Choice

Notation

Objects fixed in this chapter (formal, but lightweight).
Time $t \in \mathbb{N}_0$; agents $i \in \mathcal{A}$; actions $\{\textsc{Yes}, \textsc{No}\}$;
Refusal $R_t \in \{0,1\}$ (some culpable agent chose No at t);
Rights/consent breach $B_t \in \{0,1\}$; culpability gate $C_t \in \{0,1\}$;
realized harm $H_t \geq 0$;
Realized mercy $M_t \in \{0,1\}$; realized justice $J_t^v \in \{0,1\}$; love increment $\Delta L_t \geq 0$; freedom increment $\Delta F_t \geq 0$;
Hazard switch $\mathrm{HZ}_t \in \{0,1\}$; confirmation switch $S_t^* \in \{0,1\}$;
cumulative harm $E_{\mathrm{tot}} = \sum_{t \geq 0} H_t$;
Event-valued evaluator $\mathcal{J}$ (defined on the realized path).

7.1 Agents, actions, and live alternate possibilities

We model a finite (or countable) set of agents $\mathcal{A}$. At each time t, each agent i has at least two admissible actions, abstractly $\{\text{Yes}, \text{No}\}$, representing alignment with love/truth (Yes) or rejection (No).

Live alternate possibilities (AP)

AP are *live* at time t if for each agent i there exist at least two admissible actions that are (i) physically possible, (ii) policy-permitted, (iii) available without coercion. Formally, $\text{AP}_i(t) = 1$ if $|\mathcal{U}_i(t)| \geq 2$ with no coercive constraint eliminating a morally admissible action.

Non-coercion requires that confirmation (later S^*) is not a mind-control operation: after $S^* = 1$, freedom remains live, but certain *routes* become structurally closed.

7.2 Events on the realized path

We evaluate what *happens* (not what could have happened). At time t:

$$R_t = 1 \text{ iff } \big(B_t = 1 \text{ or } H_t \geq h_{\min}\big) \text{ and } C_t = 1,$$

where B_t flags an adjudicated rights/consent breach and C_t is the culpability gate (intentional / reckless / grossly negligent). If $C_t = 0$ (accident), record H_t for learning, but set $R_t = 0$.

Redemption-goods are event-valued:

$M_t, J_t^v \in \{0,1\}$ only when mercy/justice are *enacted* on the realized path.

7.3 Policies, trajectories, and the evaluator

A (possibly stochastic) policy π maps available information to actions for each agent. A trajectory is the realized sequence

$$\omega = \left\{(\text{YES/NO})_i(t),\ R_t,\ H_t,\ M_t,\ J_t^v,\ \Delta L_t,\ \Delta F_t,\ \text{HZ}_t,\ S_t^*\right\}_{t\geq 0}.$$

We score the single realized path via an event-valued objective:

$$\mathcal{J}(\omega) = \sum_{t\geq 0}\Big(\alpha\,\Delta L_t + \gamma\,\Delta F_t + \mu\,M_t + \nu\,J_t^v\Big) \;-\; \sum_{t\geq 0}\Big(\beta\,H_t + \kappa\,R_t\Big), \tag{7.1}$$

$$\text{s.t.}\quad \text{HZ}_t \in \{0,1\},\ S_t^* \in \{0,1\},\ E_{\text{tot}} = \sum_{t\geq 0} H_t \leq H_{\max},$$

$$\Big(\textstyle\sum_t R_t \geq 1 \ \&\ \text{remediation passes}\Big) \Rightarrow \big(S^* = 1,\ \text{HZ} = 0 \text{ thereafter}\big).$$

Weights $(\alpha, \gamma, \mu, \nu, \beta, \kappa)$ encode priorities; what matters for our results is that $\kappa > 0$ and M, J^v are event-valued.

7.4 Hazard classes and structural confirmation

Incidents are grouped into *classes* $c \in C$ sharing a causal route. For class c:

- A remediation docket records root cause, fix, tests,

owner/deadline, evidence.

- After a passed confirmation test that reenacts the prior route, set $\mathrm{HZ}(c) = 0$ and $S^*(c) = 1$ (route closed).

- Live AP remains: agents still have multiple permissible actions, but the prior hazardous route is structurally unavailable or stably unattractive.

7.5 Two feasible branches (existence)

Lemma 7.1 (Feasibility of H_0 (evil-free branch))**.** *There exists a policy $\pi^{(0)}$ under which all agents choose* YES *at all times, $R_t = 0$ for all t, so $M_t = J_t^v = 0$ for all t, and the hazard is immediately closed for all relevant classes* $(\mathrm{HZ} = 0,\ S^* = 1)$ *by construction.*

Reason. The all-YES policy is admissible (non-coercive) and entails no rights breach or harm above threshold; by the event-valued rule, redemption-goods remain 0.

Lemma 7.2 (Feasibility of H_1 (minimal-trigger redemptive branch))**.** *There exists a policy $\pi^{(1)}$ and time $t^\dagger$ such that $\sum_t R_t = 1$ with $R_{t^\dagger} = 1$ and $R_t = 0$ otherwise, followed by consent and enacted mercy/justice $(M, J^v = 1$ at some times), a passed remediation, and then structural closure* $\mathrm{HZ} = 0$ *with $S^* = 1$ thereafter.*

Reason. Take any policy that permits one culpable refusal, then mandates remediation artifacts and a confirmation test for the relevant incident class.

7.6 Grounding requirement for redemption-goods

Theorem 7.1 (Event grounding)**.** *If M_t, J_t^v are event-valued and credited only when enacted on the realized path, then on any trajectory with $R_t = 0$ for all t, one has $M_t = J_t^v = 0$ for all t. Conversely, if $\sum_t (M_t + J_t^v) \geq 1$, then $\sum_t R_t \geq 1$.*

Proof sketch. Without a realized wrong, there is nothing to pardon or rectify. Crediting M, J^v would violate the counting rule (honesty lock).

7.7 Minimal-trigger intuition (no repetition bonus)

We adopt *concave / typal* credits for redemption-goods: the first instance of a wound-type earns the kind-level credit; repeats of the same type add no new kind of value and do add cost.

Lemma 7.3 (No repetition bonus (same class))**.** *Fix an incident class c and assume concave/typal credits and $\kappa > 0$. Among policies that realize at least one R_t of type c, any policy with more than one such rejection has (weakly) lower $\mathcal{J}$ than a policy that realizes exactly one, then closes the class (*$\mathrm{HZ}(c) = 0,\ S^*(c) = 1$*).*

Reason. After the first instance, additional instances increase the cost term (κ and H_t) but add no new kind-level credit by concavity/typality; hence they cannot increase $\mathcal{J}$.

7.8 A worked micro-example (one agent, two steps)

Consider one agent and one incident class c.

$t = 0:$ YES $\Rightarrow$ $R_0 = 0,\ H_0 = 0,\ M_0 = J_0^v = 0.$

$t = 1:$ NO $\Rightarrow$ $R_1 = 1$ (culpable), $H_1 \geq h_{\min}$.

$t = 2:$ Consent $\Rightarrow$ $M_2 = 1,\ J_2^v = 1,$ remediation passes, $\mathrm{HZ}(c) = 0,\ S^*(c) = 1$

With typal/concave credits and $\kappa > 0$, any further repeats of class c after $t = 2$ would only reduce $\mathcal{J}$ relative to the one-off, repaired path.

7.9 What this chapter delivers downstream

- Precise *objects* (agents, actions, events) and *switches* (HZ, S^*) used in Parts II–III.
- Two feasible branches (Lemmas 7.1–7.2) needed for the *Two-Branch Lemma*.
- Event grounding (Theorem 7.1) used in the formal statement of MOP.
- No-repetition bonus (Lemma 7.3) used in minimal-trigger Pareto-optimality proofs.

Reader note. If you prefer the narrative first, you can skim the

definitions here, proceed to the applications, and return when proofs reference Lemmas 7.1–7.3.

Chapter 8

The State of the World

Notation

State variables (per time t; per incident class $c \in C$).

$R_t \in \{0,1\}$ (rejection event); $B_t \in \{0,1\}$ (rights/consent breach); $C_t \in \{0,1\}$ (culpability); $H_t \geq 0$ (realized harm); $M_t, J_t^y \in \{0,1\}$ (realized mercy/justice); $\Delta L_t, \Delta F_t \geq 0$ (love/freedom increments).

$\mathrm{HZ}_c(t) \in \{0,1\}$ (hazard-on for class c); $S_c^*(t) \in \{0,1\}$ (confirmation for class c); $E_{\mathrm{tot}}(t) = \sum_{\tau \leq t} H_\tau$ (cumulative harm). Global switches: $\mathrm{HZ}(t) = \max_c \mathrm{HZ}_c(t)$, $S^*(t) = \min_c S_c^*(t)$.

8.1 Incident classes and local vs. global confirmation

We partition adverse events into *incident classes* $c \in C$ that share a causal route (same vector). For each class c, the pair

$$\big(\mathrm{HZ}_c(t),\ S_c^*(t)\big) \in \{0,1\}^2$$

tracks whether that route is currently hazardous and whether it has been *structurally confirmed closed*. The *global* switches summarize system state:

$$\mathrm{HZ}(t) = \max_{c \in C} \mathrm{HZ}_c(t), \qquad S^*(t) = \min_{c \in C} S_c^*(t).$$

Thus $S^*(t) = 1$ means *all* classes are confirmed closed; $\mathrm{HZ}(t) = 0$ means *no* class is currently hazardous.

Confirmation times

For class c, define the *local confirmation time*

$$T_c^* \ := \ \inf\{t : \ S_c^*(t) = 1\}.$$

If C is finite and every class eventually confirms, the *global confirmation time* is

$$T^* \ := \ \max_{c \in C} T_c^*,$$

so that $S^*(t) = 1$ and $\mathrm{HZ}(t) = 0$ for all $t \geq T^*$.

8.2 State machine: how the world updates

We treat the world as a simple state machine updated by events and tests. For each class c, the allowed transitions are:

$$(\mathrm{HZ}_c, S_c^*) = (1,0) \xrightarrow{\text{remediation docket + passed confirmation test}} (0,1),$$

and once $(0,1)$ is reached, it is *absorbing* (no re-open) unless a new causal vector is discovered that *was not* in the scope of c (this creates a *new* class c').

Scope lock (anti-gaming)

Each class c carries a written *scope* (causal vector definition). Passing the confirmation test for c locks its scope; future failures must not be re-labeled as "the same c" unless they match the locked scope. New vectors are assigned new classes c'.

8.3 Event-to-state update rules (local)

At each time t:

(U1) Rejection trigger: $R_t = 1$ iff $\big(B_t = 1 \text{ or } H_t \geq h_{\min}\big)$ and $C_t = 1$. (8.1)

(U2) Docket obligation: if $R_t = 1$ matches class c, open/refresh a remediation

(U3) Mercy/justice credit: $M_t, J_t^v \in \{0,1\}$ only when enacted (event-valued) to

(U4) Confirmation step: upon passed reenactment test for class c, $S_c^* \leftarrow 1$, HZ (8.2)

(U5) Monotonicity: $S_c^*(t)$ is non-decreasing in t, $\quad \mathrm{HZ}_c(t)$ is non-increasing in t. (8.3)

Rules (8.1)–(8.3) implement honesty locks (event-valued credits) and structural closure.

8.4 Objective and constraints (recall)

We adopt the event-valued evaluator from Chapter 7, repeated here for reference with line breaks:

$$\mathcal{J} = \sum_{t\geq 0} \Big(\alpha\,\Delta L_t + \gamma\,\Delta F_t + \mu\,M_t + \nu\,J_t^{v}\Big) \;-\; \sum_{t\geq 0} \Big(\beta\,H_t + \kappa\,R_t\Big), \tag{8.4}$$

$$\text{s.t. } E_{\mathrm{tot}}(t) = \sum_{\tau\leq t} H_\tau \leq H_{\max}, \quad \mathrm{HZ}_c, S_c^* \in \{0,1\}, \text{ and transitions (8.2)–(8.3).}$$

Weights $(\alpha, \gamma, \mu, \nu, \beta, \kappa)$ are domain-tunable. Our results rely only on event-valued M, J^v and $\kappa > 0$.

8.5 Near-misses, drills, and measurement

- **Near-miss.** If no rights breach and $H_t < h_{\min}$, then $R_t = 0$. Still log features for prevention; use to design drills.
- **Eventized drill.** A high-fidelity reenactment of the prior causal route under stronger incentives, conducted safely. Mark a drill flag $D_t = 1$. Drills can support $S_c^* = 1$ when passed, but *do not* create M_t or J_t^v credits.

- **Evidence standards.** M_t, J_t^v require adjudication and delivered remedies; $\Delta L_t, \Delta F_t$ require verified completion (not mere intentions); H_t uses published clinical/legal scales; $h_{\min}$ is public.

8.6 Monotonicity and absorption (formal facts)

Lemma 8.1 (Monotonicity)**.** *Under update rules* (8.2)–(8.3)*, for each class c,* $S_c^*(t)$ *is non-decreasing and* $\mathrm{HZ}_c(t)$ *is non-increasing in t. In particular, once* S_c^* *flips to* 1*, it remains* 1 *thereafter; once* HZ_c *flips to* 0*, it remains* 0 *thereafter.*

Proof. Immediate from the transition design: $(0, 1)$ is absorbing for class c unless scope changes define a new class.

Lemma 8.2 (Local confirmation time exists under honest closure)**.** *If the reenacted confirmation test for class c eventually passes (with scope lock), then* $T_c^* < \infty$ *and for all* $t \geq T_c^*$ *we have* $S_c^*(t) = 1$ *and* $\mathrm{HZ}_c(t) = 0$*.*

Proof sketch. Passing the test triggers update (8.2); monotonicity (8.3) makes the state absorbing.

Theorem 8.1 (Global convergence with finitely many classes)**.** *If C is finite and every class eventually passes a confirmation test under scope lock, then* $T^* = \max_c T_c^* < \infty$ *and for all* $t \geq T^*$*,*

$$S^*(t) = 1, \qquad \mathrm{HZ}(t) = 0, \qquad R_t = 0.$$

Proof sketch. Finite maximum of the T_c^* exists; after T^*, all classes are in $(\mathrm{HZ}_c, S_c^*) = (0, 1)$, so $\mathrm{HZ}(t) = 0$ and no further R_t arise via closed vectors.

8.7 Bounded harm and redress (why repeats are dominated)

Bounded harm (A2) enforces $E_{\mathrm{tot}} \leq H_{\max}$; honesty locks forbid counting unrealized goods. With typal/concave credits for redemption-goods and $\kappa > 0$, once a class c has confirmed ($S_c^* = 1$), any further repetition of the *same* vector (i) violates the closure assumption or (ii) creates a new class c' and must pay full cost again without kind-level credit gain. This intuition is formalized in Part III.

8.8 Worked schema (per class c)

Minimal, testable lifecycle for class c:

1. **Trigger.** First $R_t = 1$ matching c; docket opened with root cause, fix, tests, owner, deadline, publication plan.

2. **Remedy.** Deliver M_t, J_t^y to the actually harmed (if applicable); implement the fix.

3. **Test.** Run a reenactment (or high-fidelity drill) of the prior route under stronger incentives; pass/fail recorded.

4. **Close.** If pass: set $S_c^* = 1,\ \mathrm{HZ}_c = 0$; publish evidence. If fail: escalate and iterate until pass or redesign scope (creating c').

8.9 What Chapter 8 delivers downstream

- A precise *state model*—local/global HZ, S^* and confirmation times T_c^*, T^*—used by the Two-Branch Lemma and by MOP.
- Monotonicity/absorption lemmas (no silent re-opening), enabling *structural* confirmation arguments in Part III.
- Measurement rules for near-misses and drills that preserve event-valued credits without rewarding unrealized goods.

Reader tip. If you prefer applications, you can skim this and proceed to later chapters; proofs in Part III will reference Lemmas 8.1–8.2 and Theorem 8.1.

Chapter 9

The Divine Objective Function

Notation

Objects fixed in this chapter. Weights $\alpha, \gamma, \mu, \nu, \beta, \kappa \geq 0$; optional discount $\delta \in (0, 1]$;
Events per t: $R_t, B_t, C_t \in \{0, 1\}$, $H_t \geq 0$, $M_t, J_t^{\nu} \in \{0, 1\}$, $\Delta L_t, \Delta F_t \geq 0$;
Incident classes $c \in C$ with switches $\mathrm{HZ}_c(t), S_c^*(t) \in \{0, 1\}$; cumulative harm $E_{\mathrm{tot}}(t) = \sum_{\tau \leq t} H_\tau$.
We score the *single realized path* (event-valued), not potentials.

9.1 Design desiderata

The evaluator should:

1. **Credit realized goods only** (event-valued, honesty lock).

2. **Subtract realized costs** (harms H_t; rejection penalty $\kappa > 0$).

3. **Respect guardrails** (non–coercion; bounded harm with redress; structural confirmation S^*).

4. **Avoid repetition bonuses** for the same wound type (typal/concave credits).

5. **Permit an evil-free branch** (coherent when $\mu = \nu = 0$) and **a minimal-trigger branch** (when $\mu, \nu > 0$).

9.2 Canonical event-valued objective (pathwise)

We adopt an additive, pathwise form. For an undiscounted horizon (bounded by guardrails):

Event-valued evaluator (canonical, undiscounted)

$$\mathcal{J} = \sum_{t \geq 0} \Big(\alpha\, \Delta L_t \;+\; \gamma\, \Delta F_t \;+\; \mu\, M_t \;+\; \nu\, J_t^{\nu} \Big) \;-\; \sum_{t \geq 0} \Big(\beta\, H_t \;+\; \kappa\, R_t \Big). \tag{9.1}$$

If you prefer an infinite horizon with convergence, use a discount

factor $\delta \in (0,1)$:

$$\mathcal{J}^{(\delta)} = \sum_{t\geq 0} \delta^t \Big(\alpha\, \Delta L_t + \gamma\, \Delta F_t + \mu\, M_t + \nu\, J_t^v\Big) \;-\; \sum_{t\geq 0} \delta^t \Big(\beta\, H_t + \kappa\, R_t\Big). \tag{9.2}$$

Only the *signs and roles* matter for our results: M_t, J_t^v are event-valued, and $\kappa > 0$.

9.3 Guardrail constraints (admissible policies)

Admissible policies π must satisfy:

Constraints & policy rule

(C1) Non-coercion: Live AP in the present; S^* cannot negate agency.

(C2) Bounded harm: $E_{\text{tot}}(t) = \sum_{\tau\leq t} H_\tau \;\leq\; H_{\max}\ (<\infty)$. (9.3)

(C3) Event-valued credits: $M_t, J_t^v \in \{0,1\}$ only when enacted on the realized p

(C4) Culpability gate: $R_t = 1$ iff $\big(B_t = 1$ or $H_t \geq h_{\min}\big)$ and $C_t = 1$. (9.4)

(C5) Confirmation rule: $\Big(\sum_t R_t \geq 1$ & remediation passes$\Big) \Rightarrow \big(S^* = 1,$ HZ (9.5)

These mirror Chapters 7–8 and will be cited in proofs.

9.4 Typal/concave redemption credits (no wound farming)

To prevent perverse incentives, redemption credits saturate by *wound type*. Let $c \in \mathcal{C}$ index incident classes (causal vectors). Let n_c be the number of realized rejections of class c on the path up to time t. Define a concave, increasing schedule $\Phi_c : \mathbb{N}_0 \to \mathbb{R}_{\geq 0}$ with $\Phi_c(0) = 0$ and $\Delta\Phi_c(n) \downarrow 0$. Then total redemption credit is bounded by

$$\sum_{t \geq 0} (\mu\, M_t + \nu\, J_t^v) \;\leq\; \sum_{c \in C} \Phi_c\big(n_c\big), \tag{9.6}$$

and in the *typal* limit, $\Phi_c(n) = \phi_c\, \mathbf{1}[n \geq 1]$ (first instance earns kind-level credit; repeats add no new kind).

9.5 Multi-agent aggregation (sums and fairness

If multiple agents $i \in \mathcal{A}$ act, we aggregate across time and agents:

$$\begin{aligned} \mathcal{J} = &\sum_{t \geq 0} \sum_{i \in \mathcal{A}} \Big(\alpha\, \Delta L_{t,i} + \gamma\, \Delta F_{t,i} + \mu\, M_{t,i} + \nu\, J_{t,i}^v\Big) \\ &- \sum_{t \geq 0} \sum_{i \in \mathcal{A}} \Big(\beta\, H_{t,i} + \kappa\, R_{t,i}\Big), \end{aligned} \tag{9.7}$$

with the same constraints class-by-class. In applications we may add fairness constraints (e.g., lower-bounds on $\sum_t \Delta F_{t,i}$ per protected group).

9.6 Two branches as optimization

The objective (9.1) makes the intuitive fork (Ch. 6) precise.

Theorem 9.1 (Two-Branch Lemma (statement only))**.** *Under* (9.1)–(9.5) *with event-valued redemption:*

- *If* $\mu = \nu = 0$*, any evil-free policy (*$R_t \equiv 0$*) attains the maximal redemption score (zero) and avoids costs; the* H_0 *branch is coherent and competitive.*

- *If* $\mu, \nu > 0$*, any policy with* $\sum_t R_t = 0$ *has* $\sum_t (M_t + J_t^{\nu}) = 0$ *(honesty lock), so achieving positive redemption score requires* $\sum_t R_t \geq 1$.

Proof is deferred to Chapter 10 (we already proved the grounding piece in Theorem 7.1).

9.7 Minimal-trigger dominance (role of κ and concavity)

When $\kappa > 0$ and (9.6) holds, repeating the same wound type cannot raise $\mathcal{J}$ relative to a one-off, repaired instance followed by closure (HZ = 0, $S^* = 1$). Chapter 11 formalizes this as Pareto-optimality of the *minimal-trigger* arc.

9.8 Capacity-valued comparator (why opinions differ)

For contrast, a capacity-valued evaluator $\mathcal{J}^{\dagger}$ that credits "could-be mercy/justice" without realized wrong would replace M_t, J_t^{ν} by *abilities* or *opportunities*. Then an evil-free policy could earn redemption credit, and our conditional necessity would not apply. The disagreement is not about algebra but about *what counts*.

9.9 Well-posedness (existence & finiteness)

If either (i) $\delta < 1$ in (9.2) or (ii) $E_{\text{tot}} \leq H_{\max}$ and only finitely many classes confirm (Ch. 8), then $\mathcal{J}$ is finite on admissible policies. Feasible policies exist by Lemmas 7.1–7.2 (the H_0 and H_1 constructions).

9.10 Worked toy program (one class)

One incident class c, typal credits $\Phi_c(n) = \phi\,\mathbf{1}[n \geq 1]$, and two policies:

$$\pi^{(0)}:\ R_t \equiv 0, \quad \mathcal{J}(\pi^{(0)}) = \sum_t (\alpha \Delta L_t + \gamma \Delta F_t) - \sum_t \beta H_t.$$

$$\pi^{(1)}:\ \exists\, t^{\dagger}: R_{t^{\dagger}} = 1,\ \text{remedy} \Rightarrow M = J^{\nu} = 1,\ \text{HZ} = 0,\ S^{*} = 1.$$

Then

$$\mathcal{J}(\pi^{(1)}) - \mathcal{J}(\pi^{(0)}) \;=\; \phi \;-\; \kappa \;-\; \sum_{t \in \text{repair}} \beta H_t \;+\; \Delta(\alpha L + \gamma F).$$

If ϕ (the single kind-level redemption credit) exceeds the intrinsic penalty plus repair costs (net of love/freedom gains), $\pi^{(1)}$ dominates; additional repeats of c can only subtract $\kappa + \beta H$ without increasing ϕ.

9.11 What Chapter 9 gives the proofs

- A canonical evaluator (9.1)/(9.2) with guardrails (9.3)–(9.5).
- The formal setup for the Two-Branch Lemma and MOP (Ch. 10).
- The exact role of typal/concave credits and $\kappa > 0$ used in minimal-trigger optimality (Ch. 11).

Chapter 10

Two Feasible Paths — Lemmas and Proofs

Notation

Standing assumptions and objects (from Chs. 7–9).
(A1) Non-coercion (live AP); (A2) Bounded harm with redress; (A3) Honesty locks (event-valued credits, no repetition bonus); (A4) Structural confirmation.
Events: $R_t, B_t, C_t \in \{0,1\}$, $H_t \geq 0$, $M_t, J_t^v \in \{0,1\}$, $\Delta L_t, \Delta F_t \geq 0$; switches $\mathrm{HZ}_c(t), S_c^*(t) \in \{0,1\}$; cumulative harm $E_{\mathrm{tot}} = \sum_{\tau \leq t} H_\tau$.
Evaluator (pathwise): $\mathcal{J}$ as in (9.1)–(9.2).

10.1 The two feasible branches (existence and structure)

Lemma 10.1 (Evil-free branch H_0)**.** *There exists an admissible policy $\pi^{(0)}$ such that $R_t \equiv 0$ for all t; consequently $M_t = J_t^{\nu} \equiv 0$, hazard is closed by construction and $S^* = 1$. Live AP is preserved (A1).*

Proof. Take the all-YES policy (Ch. 7, Lemma 7.1). Since no rights breach or $H_t \geq h_{\min}$ occurs, $R_t = 0$ by the gate (9.4). By the event-valued rule, $M_t = J_t^{\nu} = 0$ (no wrong to remedy). One may set HZ = 0 and $S^* = 1$ trivially for all incident classes without violating A1 (multiple admissible actions remain).

Lemma 10.2 (Minimal-trigger redemptive branch H_1)**.** *There exists an admissible policy $\pi^{(1)}$ under which exactly one $t^{\dagger}$ has $R_{t^{\dagger}} = 1$ and $R_t = 0$ otherwise; thereafter consent is given, M, J^{ν} are enacted (at some times), remediation passes, the relevant class is closed $\big(\mathrm{HZ} = 0\big)$, and $S^* = 1$ with live AP preserved.*

Proof. Ch. 7, Lemma 7.2 constructs such a policy: permit a single culpable refusal, then require remediation and a passed confirmation test (Ch. 8). Monotonicity gives absorption $(\mathrm{HZ}, S^*) = (0, 1)$ afterwards.

10.2 Event grounding (needed for the fork)

Theorem 10.1 (Event grounding, restated)**.** *If M_t, J_t^{ν} are event-valued and credited only when enacted on the realized path, then on*

any trajectory with $R_t = 0$ for all t, we must have $M_t = J_t^v = 0$ for all t. Conversely, if $\sum_t (M_t + J_t^v) \geq 1$, then $\sum_t R_t \geq 1$.

Proof. This is Theorem 7.1 in Ch. 7. No realized wrong $\Rightarrow$ nothing to pardon or rectify; crediting would violate (A3).

10.3 The Two-Branch Lemma (formal statement)

Theorem 10.2 (Two-Branch Lemma)**.** *Under (A1)–(A4),* (9.1)–(9.5)*, and event-valued M, J^v:*

1. *(*Evil-free feasibility*) The branch H_0 in Lemma 10.1 is feasible. It yields $\sum_t (M_t + J_t^v) = 0$.*
2. *(*Redemptive feasibility*) The branch H_1 in Lemma 10.2 is feasible. It yields $\sum_t (M_t + J_t^v) \geq 1$ with exactly one realized $R_t = 1$.*
3. *(*Exclusivity in redemption*) If $\sum_t (M_t + J_t^v) \geq 1$, then H_0 is impossible (by Theorem 10.1); some $R_t = 1$ must occur.*

Proof. Items (1) and (2) are Lemmas 10.1–10.2. Item (3) follows from Theorem 10.1.

10.4 Margan's Optimization Paradox (MOP)

We now make precise the informal triad: *eternal love*, *live AP at every moment*, and *zero rejection at every moment*. The key is what "eternal love" demands in an *event-valued* ledger.

Eternal love target (event-valued reading)

Say the *eternal love target* holds if, along the realized path,

$$\sum_{t\geq 0}\big(\Delta L_t\big) \text{ is maximal subject to (A1)–(A4)} \quad \text{and} \quad \sum_{t\geq 0}(M_t{+}J_t^{v}) \geq \rho > 0,$$

i.e., love-goods are realized in the limit and redemption-goods are not identically zero. (The threshold $\rho > 0$ can be any positive constant.)

Theorem 10.3 (MOP — conditional incompatibility)**.** *Assume (A1)–(A4), event-valued M, J^{v}, and $\mu, \nu > 0$ in* (9.1)*. The triad*

(i) eternal love target + *(ii) live AP at every t* + *(iii) $R_t = 0\,\forall t$*

is jointly inconsistent. In particular, if (ii) and (iii) hold, then (i) fails in its redemption component: $\sum_t (M_t + J_t^{v}) = 0$.

Proof. If $R_t = 0$ for all t, Theorem 10.1 gives $M_t = J_t^{v} = 0$ for all t, contradicting the nonzero redemption requirement in the *eternal love target*. Live AP (A1) is compatible with either branch and does not rescue the contradiction.

Corollary (When the triad *is*, breakable)

If one adopts a capacity-valued evaluator $\mathcal{J}^{\dagger}$ in which "could-be mercy/justice" are credited without realized wrong *or* one sets $\mu = \nu = 0$ (no value for event redemption-goods), then $R_t \equiv 0$ with live AP is coherent and competitive (the H_0 branch).

10.5 Minimal-trigger necessity (when redemption matters)

Theorem 10.4 (Minimal-trigger necessity)**.** *Under (A1)–(A4) and event-valued M, J^{ν} with $\mu, \nu > 0$, any policy achieving $\sum_t (M_t + J_t^{\nu}) \geq 1$ must realize $\sum_t R_t \geq 1$. Among such policies, if credits for a given incident class c are concave/typal and $\kappa > 0$, then realizing more than one R_t of class c cannot increase $\mathcal{J}$ relative to one-off + closure* $\big(\mathrm{HZ}(c) = 0,\ S^*(c) = 1\big)$.

Proof. The first claim is Theorem 10.1. The second follows from the no-repetition-bonus Lemma 7.3 (Ch. 7) together with the cost terms in (9.1).

10.6 What MOP is *not* claiming

- **Not a denial of evil-free coherence.** H_0 is coherent and admissible when $\mu = \nu = 0$ or under capacity-valued evaluation.

- **Not coercion.** $S^* = 1$ is a structural closure of a route, not mind control; AP remain live (A1).

- **Not a license for harm.** With $\kappa > 0$, bounded harm (A2), and concave credits, repetition is dominated; the evaluator penalizes realized costs and compels structural closure.

10.7 Roadmap to Chapter 11

Chapter 11 upgrades the qualitative necessity above to a *Pareto-optimality* statement: under typal/concave credits, $\kappa > 0$, and the guardrails, the *minimal-trigger* arc (one realized refusal of a class, then remedy, closure, confirmation) weakly dominates any policy with extra repeats of that class.

Chapter 11

Pareto-Optimality of the Minimal-Trigger Arc

Notation

Standing assumptions. (A1) Non-coercion (live AP); (A2) Bounded harm with redress ($E_{\text{tot}} \le H_{\max}$); (A3) Honesty locks (event-valued credits; no repetition bonus); (A4) Structural confirmation (HZ$\downarrow$ 0, $S^{*}\uparrow$ 1 after passed test).
Evaluator $\mathcal{J}$ as in (9.1) or (9.2); redemption credits concave/typal by class $c \in C$ as in (9.6); $\kappa > 0$.

11.1 What we prove

Intuitively: once a *type* of wound has been realized and repaired, repeating the *same* type cannot improve the event-valued evaluator

when costs are counted and credits are concave by type. We make this precise in two steps:

1. **Single-class dominance.** For a fixed incident class c, among all admissible policies that realize $n \geq 1$ rejections of class c, $\mathcal{J}$ is maximized at $n = 1$ followed by closure ($\mathrm{HZ}(c) = 0,\ S^*(c) = 1$).

2. **Multi-class composition.** If classes are additively separable in credits/costs (no positive cross-credit from repeating a closed class), then applying step (1) to every class is Pareto-improving; any policy with repeats is (weakly) dominated by one that has at most one realized refusal per class.

We also give *strict* dominance conditions (the inequality is $>$, not just $\geq$) and note that discounting $\delta < 1$ strengthens the results.

11.2 Definitions and a simple exchange argument

Dominance and equivalence (restricted to a class)

Fix a class c. Two admissible policies π, π' are *equivalent outside* c if their realized paths agree on all variables and classes $c' \neq c$ (same $\Delta L_t, \Delta F_t, H_t, M_t, J_t^y$ for $c' \neq c$; same $\mathrm{HZ}(c'), S^*(c')$ trajectories). We say π' *dominates* π on c if π' is equivalent outside c and $\mathcal{J}(\pi') \geq \mathcal{J}(\pi)$ with strict $>$ in the strict case.

Lemma 11.1 (Exchange step (remove one repeat of class c))**.** *Assume typal/concave credit* (9.6) *by class c and $\kappa > 0$. Let π realize $n \geq 2$*

rejections of class c at times $t_1 < \cdots < t_n$. There exists an admissible π' equivalent outside c such that (i) π' realizes only the first of those rejections ($n' = 1$), (ii) implements remedy and closure immediately after t_1, and (iii) $\mathcal{J}(\pi') \geq \mathcal{J}(\pi)$, with strict $>$ if any added cost beyond the first occurs in π (extra κ or extra H_t).

Proof sketch. Concavity/typality implies the total redemption credit attainable from class c is bounded by $\Phi_c(n)$ with $\Delta\Phi_c(n) \downarrow 0$ and, in the typal limit, $\Phi_c(n) = \phi_c \mathbf{1}[n \geq 1]$. Hence $\Phi_c(n) = \Phi_c(1)$ for $n \geq 1$ in the typal case and $\Phi_c(n) \leq \Phi_c(1) + \sum_{k=2}^{n} \Delta\Phi_c(k)$ with $\Delta\Phi_c(k) \to 0$ in the concave case. Meanwhile, each additional realized rejection contributes at least $\kappa > 0$ and typically additional H_t to the cost sum. Construct π' by copying π up to t_1, then performing the same remediation and confirmation sequence that π eventually performs after t_n, but *immediately* after t_1. Since credits do not increase by repeating the same class while costs do, $\mathcal{J}(\pi') \geq \mathcal{J}(\pi)$, with strict inequality if any positive incremental cost is present after t_1 in π.

11.3 Single-class Pareto-optimality

Theorem 11.1 (Single-class minimal-trigger dominance)**.** *Fix class c and assume* (9.6) *with $\kappa > 0$ and (A1)–(A4). Among admissible policies that realize at least one R_t of class c, any policy with $n \geq 2$ such rejections is dominated (on c) by a policy with $n = 1$ followed by closure (*$\mathrm{HZ}(c) = 0$, $S^*(c) = 1$*). If any positive incremental cost occurs after the first (κ or H_t), the dominance is strict.*

Proof. Iteratively apply Lemma 11.1 to remove repeats until $n =$

1. Each step weakly improves $\mathcal{J}$ and preserves feasibility (A1–A4). If any step removes a strictly positive incremental cost, strict improvement occurs.

11.4 Multi-class composition (no cross-credit from repeats)

Additive separability by class

Credits and costs aggregate by class without positive cross-credit for repeating a closed class:

$$\sum_{t\geq 0}(\mu M_t+\nu J_t^y) = \sum_{c\in C}\sum_{t\geq 0}(\mu M_{t,c}+\nu J_{t,c}^y), \quad \sum_{t\geq 0}(\beta H_t+\kappa R_t) = \sum_{c\in C}\sum_{t\geq 0}(\beta H_{t,c}+\kappa R_{t,c})$$

Moreover, once $S^*(c) = 1$ with $\mathrm{HZ}(c) = 0$, repeating c cannot increase any other class's redemption credit.

Theorem 11.2 (Global minimal-trigger dominance)**.** *Under additive separability and assumptions above, any admissible policy π is (weakly) dominated by a policy $\tilde{\pi}$ that, for each class c, realizes at most one R_t of type c and then confirms closure for c. If any class has a strictly positive incremental cost beyond the first instance, the dominance is strict.*

Proof. Apply Theorem 11.1 to each class c while holding the others fixed (equivalence outside c). Because cross-credits do not increase by repeating a closed class, each replacement weakly improves $\mathcal{J}$. If any replacement removes strictly positive incremental cost, we obtain strict improvement.

11.5 Discounting strengthens the result

Theorem 11.3 (Discounted preference for earlier closure)**.** *If $\delta \in (0, 1)$ in* (9.2)*, then for any class c and any pair of admissible policies that both realize exactly one R_t of class c and the same total costs/credits, the policy that closes earlier (smaller confirmation time T_c^*) attains weakly higher $\mathcal{J}^{(\delta)}$, with strict inequality if any positive flow (cost or credit) is time-shifted.*

Proof. Standard discounting: earlier positive flows and later negative flows both increase present value; earlier closure reduces expected future costs and speeds any time-limited benefits (e.g., restored $\Delta L_t, \Delta F_t$).

11.6 Strictness conditions (when > is guaranteed)

Strict dominance arises under any of the following:

- **Intrinsic penalty.** $\kappa > 0$ and at least one extra rejection beyond the first occurs.
- **Repair costs.** Any positive harm H_t or remediation burden strictly increases with repeats.
- **Typal credits.** $\Phi_c(n) = \phi_c \, \mathbf{1}[n \geq 1]$ (no extra credit beyond the first).
- **Discounting.** $\delta < 1$ and closure times differ (Theorem 11.3).

11.7 What if classes interact? (limits and variants)

If classes interact so that repeating c could *increase* redemption credit in another class c' (a cross-credit), our dominance result requires an additional honesty lock: no cross-credit unless a *new* class c' would independently realize credit without reusing c's route. In practice we enforce this with *scope lock* (Chapter 8): once c is closed, its route cannot be relabeled to count for another class unless the causal vector truly differs (creating a new class c').

11.8 Minimal-trigger doctrine (policy form)

Minimal-trigger doctrine (per class c)

1. If $R_t = 1$ of type c occurs, open a remediation docket immediately (root cause, fix, tests, owner, deadline, publish plan).
2. Deliver remedy to the actually harmed (event-valued): enact M, J^v where applicable.
3. Run a reenactment confirmation test under stronger incentives; on pass, set $S^*(c) = 1$ and $\mathrm{HZ}(c) = 0$.
4. Do *not* count repeats of class c as new kinds of value; they incur costs and are dominated by the one-off, repaired route.

11.9 Consequences for design and governance

- **No wound farming.** Systems must not reward repeated "redemption stories" of the same kind.
- **Close classes.** After the first failure and remedy, the question is not "who to blame next" but "has the route been closed and proven closed?".
- **Early closure preferred.** With discounting or practical risk, earlier confirmation strictly improves outcomes.

11.10 What Chapter 11 provides downstream

We have upgraded the necessity result (Chapter 10) to a *Pareto-optimality* statement: one realized refusal *per class* followed by remedy and structural confirmation weakly dominates any policy that repeats the same class, with strict dominance under mild conditions. Applications in Part IV will adopt this as a design rule across AI safety, healthcare, justice, and policy.

Part III

The Optimal Arc

Chapter 12

Mapping the Narrative: From Permission to Confirmation

Notation

Aim. Connect the formal arc *permission* → *rectification* → *confirmation* to a theological narrative without changing the evaluator. We keep everything *event-valued*: we only count what history enacts. Symbols used: R_t (rejection), B_t (rights/consent breach), C_t (culpability), H_t (harm), M_t (mercy), J_t^y (justice), HZ (hazard on/off), S^* (confirmation), $\mathcal{J}$ (evaluator).

12.1 Two branches, retold theologically

We interpret the two feasible branches (Chs. 7–10) in narrative terms.

Evil-free branch H_0. Creatures freely choose YES from the outset; no realized wrong occurs, so $M_t = J_t^v = 0$ forever by the honesty lock. This branch is coherent if one either (i) sets $\mu = \nu = 0$ (does not value *realized* redemption-goods) or (ii) adopts a capacity-valued comparator $\mathcal{J}^\dagger$ that counts "could-be mercy/justice" without events.

Redemptive branch H_1 (minimal trigger). A *single* realized rejection occurs ($\sum_t R_t = 1$), followed by consent, enacted mercy and justice to the actually harmed (M, J^v on history), structural hazard closure (HZ = 0), and confirmation ($S^* = 1$) with live AP preserved. This branch exists whenever $\mu, \nu > 0$ (you value *realized* redemption-goods) and guardrails hold.

12.2 A canonical mapping of the big story

We sketch one natural alignment between the formal arc and a classical Christian storyline.[1]

[1] Readers from other traditions can re-map using the same evaluator (see Ch. 27 on plural alternatives).

Formal stage	**Narrative element**	**Event-valued evidence (what is counted)**
Permission of refusal	Fall of rational agents (angelic/human)	Adjudicated rejection $R_t = 1$ (rights/consent breach or harm $\geq h_{\min}$ with culpability $C_t = 1$); realized harms H_t accumulate
Rectification (mercy/justice)	Atonement/forgiveness; judgment/righteousness	$M_t = 1$ when pardon/restoration is enacted to those actually harmed; $J_t^y = 1$ when rectification/accountability is enacted
Hazard reset	New-covenant / new-heart promises; transformed practices	Passed reenactment test for the prior causal route $\Rightarrow$ HZ = 0 (class closed); scope locked (no relabeling)
Confirmation	Eschatological security / perseverance	$S^* = 1$ with live AP: freedom remains, but the prior hazardous route is structurally unavailable/stably unattractive

12.3 Mercy and justice as *event* goods

In our ledger, mercy and justice are not moods or potentials; they are *events*. We therefore require a *grounding wrong* on the realized path (Ch. 7). The *Margan Optimization Paradox* (Ch. 10) follows: under event-valued credits and $\mu, \nu > 0$, the triad (eternal love target + live AP at every moment + zero rejection at every moment) is inconsistent.

Co-witness principle (mercy and justice together)

A rectification is complete only when mercy to the harmed and justice toward the wrong are both enacted *in history*. Formally, a docket for a class c must include both $M = 1$ (restorative enactment) and $J^{\nu} = 1$ (rectificatory/forensic enactment) before closure.

12.4 Sanctification without coercion (character ≠ mind control)

Structural confirmation (restated for the narrative)

After remediation passes for a class c, we set $S^{*}(c) = 1$ and force $\mathrm{HZ}(c) = 0$. Agents retain live alternate possibilities; nevertheless, the exact prior causal route is closed by design and by formed character (stable aversion to the old vector).

This preserves A1 (non–coercion) while explaining how an eternal future can be secure: not by lobotomy, but by structure and

character.

12.5 Why one trigger is enough (and more is worse)

With typal/concave credits and $\kappa > 0$, repeating the *same* wound type can only lower the evaluator (Ch. 11). The narrative upshot is stark:

> If one trigger suffices to realize all kind-level redemption goods, then *minimize* the realized refusal and close the route; do not farm wounds for "more story."

12.6 Micro-narratives that instantiate the arc

The big pattern repeats in smaller ones:

- **Personal failure → repentance → stability.** A denial or betrayal ($R_t = 1$) followed by confession and restitution (M, J^v), then a formed character; the old route becomes stably unattractive (local $S^* = 1$).
- **Community breach → discipline → restoration.** Safeguarding failure or injustice (rights breach) addressed by transparent remedy and structural fixes; confirmation via reenactment drills; route closed for the community.

These micro-cases provide empirical handles: the same doctrine is testable at human scale.

12.7 Objections and replies (narrative form)

Why allow any rejection at all? Because under the event-valued reading of redemption-goods ($\mu, \nu > 0$), some realized wrong is required to *ground* mercy/justice. The honesty lock forbids credit without event.

Why not create creatures confirmed from the start? That would sacrifice either live AP (A1) or the redemption component of the eternal love target. Under capacity-valued $\mathcal{J}^{\dagger}$, "confirmed-from-start" is coherent; under event-valued $\mathcal{J}$ with $\mu, \nu > 0$, H_0 cannot realize redemption-goods.

Does this justify evil? No. With $\kappa > 0$, bounded harm (A2), and typal/concave credits (A3), repetition is dominated, and closure is mandatory (A4). The doctrine is *anti*-license: it compels the *least* realized wrong that suffices for the goods you *actually* value.

12.8 The timeline written thin (fits on one page)

$$\mathbf{H}_1\text{:}\quad \underbrace{R=1}_{\text{grounding wrong}} \Rightarrow \underbrace{\text{consent}}_{\text{turn}} \Rightarrow \underbrace{M{+}J^{\nu}}_{\text{enacted repair}} \Rightarrow \underbrace{\text{passed reenactment}}_{\text{evidence}} \Rightarrow \underbrace{\text{HZ}=0}_{\text{route closed}} \Rightarrow \underbrace{S^{*}=1}_{\text{stable future; AP live}}$$

12.9 Eschatological T^* (global confirmation time)

If only finitely many incident classes exist (Ch. 8) and each passes a confirmation test under scope lock, the global time

$$T^* = \max_{c \in C} T_c^*$$

is finite (Theorem 8.3). Theologically, this models a consummation: all known hazard vectors closed; freedom live; rejection stably absent.

12.10 What Chapter 12 delivers

- A consistent mapping from the formal arc to a familiar narrative without changing the evaluator or guardrails.
- A clear role for *events*: mercy/justice *count* only when enacted, compelling the minimal-trigger doctrine.
- Empirical hooks: micro-narratives and community processes that let readers test the claims at human scale.

Chapter 13

Autonomous Vehicles: Minimal-Trigger Safety Doctrine

Notation

Domain mapping (AV). An *event* is a logged traffic interaction with adjudication potential.

$R_t \in \{0,1\}$: rejection event (culpable safety rule breach or harm $\geq h_{\min}$); $B_t \in \{0,1\}$: traffic/consent breach; $C_t \in \{0,1\}$: culpability gate; $H_t \geq 0$: realized harm (severity scale).

$M_t, J_t^v \in \{0,1\}$: enacted mercy/justice (compensation/restoration; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: safety-culture and freedom increments;

Incident classes $c \in C$: causal routes (e.g., "left-turn across

path" LTAP, "occluded pedestrian at night," "phantom braking").
Switches: $\mathrm{HZ}_c(t) \in \{0,1\}$ (hazard on/off per class), $S_c^*(t) \in \{0,1\}$ (confirmation per class).
Evaluator: $\mathcal{J}$ as in Ch. 9, event-valued on the realized path.

13.1 Why AVs fit the event-valued approach

Autonomous driving already logs high-fidelity timelines (sensors, decisions, actuator commands). This makes AVs ideal for *event-valued* evaluation: we can tie credits/debits to what *actually* happened, not what might have happened. The doctrine we apply:

> One realized failure of a given causal kind *triggers* docketed remediation and a *confirmation test* that proves the route is closed ($\mathrm{HZ}(c) = 0,\ S^*(c) = 1$). Repeating the same kind brings costs but no new kind-level credit.

13.2 Operational semantics for AV variables

Rejection R_t. Set $R_t = 1$ iff a culpable breach occurs:

$$R_t = 1 \iff \big(B_t = 1 \,(\text{traffic/consent breach}) \text{ or } H_t \geq h_{\min}\big) \text{ and } C_t = 1.$$

$C_t = 1$ when evidence shows intent/recklessness/gross negligence by an accountable agent (model operator/organization) relative to a published safety case; else $C_t = 0$ (accident; record H_t but do not

count rejection).

Harm H_t. Use a severity scale (e.g., property-only $< h_{\min}$; minor injury; serious injury; fatality), with $h_{\min}$ published per jurisdiction.

Mercy/Justice M_t, J_t^v. Event-valued only: $M_t = 1$ when compensation/restoration is enacted to the actually harmed; $J_t^v = 1$ when rectifying/accountability actions are enacted (e.g., recall, public notice, sanction, or verifiable process correction).

13.3 Incident classes (examples)

Typical AV classes c (illustrative, not exhaustive):

- **LTAP** (Left-turn across path of oncoming vulnerable road user).
- **Occlusion/low-light pedestrian** (missed detection due to occlusion + low illumination).
- **Phantom braking** (false-positive obstacle $\rightarrow$ abrupt decel with rear-end risk).
- **Map mislocalization at merge** (lane inference error under construction).
- **Unprotected left at multi-agent junction** (gap acceptance misjudgment).

Each class must have a *scope lock*: what sensor stack, perception module, scene preconditions, and planner states define the causal route.

13.4 AV evaluator (domain view; line-broken to fit)

$$\begin{aligned}\mathcal{J}_{\mathrm{AV}} = &\sum_{t\geq 0}\Big(\alpha\,\Delta L_t + \gamma\,\Delta F_t + \mu\,M_t + \nu\,J_t^{v}\Big)\\ &-\sum_{t\geq 0}\Big(\beta\,H_t + \kappa\,R_t\Big), \qquad (13.1)\\ \text{s.t.}\quad &E_{\mathrm{tot}} = \sum_{\tau\leq t} H_\tau \leq H_{\max},\ \ \mathrm{HZ}_c, S_c^{*} \in \{0,1\},\ \ \text{updates per Ch. 8.}\end{aligned}$$

Interpretation: ΔL_t counts verified safety-culture improvements delivered (not proposed), ΔF_t counts restored freedoms (e.g., re-opened ODD after proof of safety), and (M_t, J_t^{v}) are *enacted* remedies. $\kappa > 0$ and concave/typal redemption credits prevent "wound farming."

13.5 Minimal-trigger doctrine for AVs (policy form)

Per class c

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket* with: root cause (perception/planning/controls), fix, tests, owner & deadline,

publication plan.

2. **Remedy:** Enact M_t (compensation/restoration) and J_t^v (rectification/accountability).

3. **Confirm:** Run reenactment tests: high-fidelity sim $\rightarrow$ closed-course $\rightarrow$ monitored on-road, with stronger incentives than the original failure; on pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats of class c add cost but no new kind-level credit; they are dominated by the one-off, repaired route.

13.6 Evidence pack (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Police/medical reports; calibrated severity scale; time-stamped telemetry
B_t (breach)	Reconstruction showing rule violation (speed, right-of-way, signal phase)
C_t (culpability)	Safety-case gap (foreseeable, preventable with due care); governance logs
M_t (mercy)	Verified compensation/restoration delivered to harmed parties
J_t^v (justice)	Recall/patch issued; accountability enacted; regulatory filing
ΔL_t	Completed safety practice: new checklists, gating, training with audit trail
ΔF_t	Re-opened ODD or restored capability after proof; due-process logs
S_c^*	Passed reenactment(s) with stronger incentives; publication of test artefacts

13.7 Confirmation tests (design pattern)

Step 1: Simulation. Recreate the exact causal vector (same lighting, occlusion, approach speeds); vary nuisance parameters; pass/fail on collision/near-miss metrics.

Step 2: Closed course. Instrumented actors; identical geometry;

pass at worst-case envelopes.

Step 3: Monitored on-road. Shadow/limited ODD with watchdogs; publish incident-free mileage for the class.

Scope lock. Freeze the class definition; if a new vector appears, create c' (no relabeling to harvest credit).

13.8 Micro-vignette (worked example)

Class: Occluded pedestrian at night.

t=0 (trigger): AV turns right; occluded pedestrian steps from behind a van; impact causes injury ($H_t \geq h_{\min}$). Investigators find perception pipeline failed under low-light + partial occlusion; $C_t = 1$ via safety-case gap $\Rightarrow R_t = 1$.

t=1 (remedy): Company compensates victim ($M_t = 1$), issues rectifying actions ($J_t^v = 1$): new low-light dataset, retrained detector, planner rule for conservative creep, crosswalk illumination requirement.

t=2 (confirm): Sim suite shows no regression; closed-course night tests pass with mannequins at varied speeds; monitored on-road phase produces 0 recurrences over a predefined mileage budget. Set $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

Result: Any further failure of the same vector would add cost (κ and H_t) with no new kind-level credit; per Ch. 11, the minimal-trigger route dominates.

13.9 Metrics & dashboards (event-ledger view)

- **Harm:** KSI (killed/seriously injured) count and rate; time-to-remedy; recurrence rate by class.
- **Closure:** Number of classes with $S^* = 1$; median T_c^*; backlog of open dockets.
- **Culture:** Completed checklists, drills run, audit pass-rate (ΔL_t proxies).
- **Freedom:** ODD restored after proof; tool-gating removed after closure (ΔF_t).

13.10 Tool gating and staged deployment

After a trigger in class c, impose *gates* until confirmation: reduce speed caps, shrink ODD, require human supervisor approval for edge scenarios, enable redundant sensing, and add conservative planner settings. Remove gates only after $S^*(c) = 1$.

13.11 Why repeats are dominated (AV intuition)

With typal/concave credits and $\kappa > 0$, repeating a closed class adds no new kind-level value and incurs fresh cost (downtime, recalls, sanctions, H_t). Discounting ($\delta < 1$) strengthens the preference for *earlier* closure (Ch. 11).

13.12 Anti-gaming: suppression and scope creep

- **Suppression.** All R_t candidates must be logged; whistleblower channels; penalties for under-reporting. Auditors sample raw telemetry against incident registers.
- **Scope creep.** Lock class scope before testing; if a later failure differs causally, open c' rather than relabeling.

13.13 One-page audit checklist (drop-in)

AV Incident → Confirmation Checklist

Trigger captured? Event time, location, actors, telemetry hash

Gate set? ODD shrink, speed cap, supervisor approval, watchdogs

Docket opened? Root cause, fix, tests, owner, deadline, publish plan

Remedy enacted? M_t delivered to harmed; J_t^y accountability done

Tests passed? Sim ✓ Closed-course ✓ On-road ✓ (evidence linked)

Scope locked? Class definition frozen; new vectors registered as c'

Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window documented

Gates removed? Only after closure; rationale filed

13.14 Limits and open problems

- **Attribution.** Assigning culpability C_t in multi-actor mixes (AV + human drivers + infrastructure) can be non-trivial.
- **Rare hazards.** Low base rates make T_c^* estimation slow; use high-fidelity drills to complement scarce events.
- **Shifting ODD.** ODD changes can create new classes (c'); honest scoping is essential.
- **Spec drift.** Model updates change distributions; keep regression suites synchronized with class scopes.

13.15 What this chapter contributes

A complete translation of the minimal-trigger doctrine into an AV safety program: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and evidence standards that make repetition of the same class strictly dominated once costs are counted. This is *event-valued* safety: promises do not count; enacted fixes do.

Chapter 14

Clinical AI: Consent, Harm, and Confirmation in Care

Notation

Domain mapping (healthcare/clinical AI).
An *event* is a documented clinical interaction or system action with adjudication potential.
$R_t \in \{0,1\}$: **rejection** (culpable rights/consent breach or harm $\geq h_{\min}$); $B_t \in \{0,1\}$: rights/consent breach (e.g., GDPR/PHI, treatment without consent); $C_t \in \{0,1\}$: culpability gate (intent/reckless/gross negligence); $H_t \geq 0$: realized clinical harm (sentinel/never-event scales).
$M_t, J_t^v \in \{0,1\}$: enacted mercy/justice (disclosure & apology, restitution; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: love/freedom increments (safeguarding done, rights exercised).

Incident classes $c \in C$: causal routes (e.g., med error—LASA[a], sepsis miss, data leak, mis-triage).
Switches: $\mathrm{HZ}_c(t) \in \{0,1\}$ (hazard on/off), $S_c^*(t) \in \{0,1\}$ (confirmation).
Evaluator $\mathcal{J}$ is event-valued as in Ch. 9.

[a] Look-Alike/Sound-Alike.

14.1 Why clinical AI benefits from an event-valued ledger

Clinical AI systems operate where harms are concrete and stakes high. Traditional governance often tallies *policies and intentions*; we insist on *events*: what actually happened to patients, what was repaired, and whether the prior causal route is *structurally* closed ($\mathrm{HZ}=0$, $S^*=1$). This avoids two failures: (i) over-crediting proposed fixes, and (ii) under-penalizing repeat vectors.

14.2 Operational semantics (how variables instantiate)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \text{ (rights/consent breach) or } H_t \geq h_{\min}\big) \text{ and } C_t = 1.$$

$C_t = 1$ when an accountable clinician/organization had adequate knowledge & freedom and crossed at least one fault threshold (intent,

recklessness, gross negligence). If $C_t = 0$ (accident), record H_t but do not set $R_t = 1$.

Harm H_t. Use a published clinical scale (e.g., near-miss $< h_{\min}$, minor harm, moderate, severe, death). Publish $h_{\min}$ per service line.

Mercy/justice M_t, J_t^v (event-valued only). $M_t = 1$ when disclosure/apology/restoration is enacted to the harmed; $J_t^v = 1$ when rectification/accountability is *enacted* (policy correction, sanction where appropriate, restitution delivered).

Near-misses and drills. Near-misses ($H_t < h_{\min}$ and no B_t) are *not* $R_t = 1$ but should spawn *eventized drills* ($D_t = 1$) that reenact dangerous scenarios safely; drills support confirmation but do not mint M, J^v.

14.3 Typical clinical AI incident classes (c)

- **LASA medication error via AI order-set suggestion.**
- **Sepsis mis-triage** (risk model under- or over-scores due to shift).
- **Imaging misclassification** (distribution drift; under-served subgroup).
- **Data leak / privacy breach** (PHI exposed via tool integration).
- **Unsafe override pattern** (AI recommendation routinely bypassed without reason logging).

- **Handoff failure** (AI summary omits critical info; adverse event ensues).

Each class carries a written *scope lock*: inputs, model version, thresholding logic, workflow, and clinical preconditions.

14.4 Evaluator and constraints (domain view; broken to fit)

$$\mathcal{J}_{\text{clin}} = \sum_{t \geq 0} \Big(\alpha\, \Delta L_t + \gamma\, \Delta F_t + \mu\, M_t + \nu\, J_t^{v} \Big) - \sum_{t \geq 0} \Big(\beta\, H_t + \kappa\, R_t \Big), \tag{14.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau \leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^{*} \in \{0,1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t counts *delivered* safeguarding/quality actions; ΔF_t counts *actually exercised* rights (consent recorded, second opinions accessible); (M_t, J_t^{v}) are enacted remedies; $\kappa > 0$ penalizes rejections.

14.5 Minimal-trigger doctrine for clinical AI

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (data/threshold/workflow), fix, tests, owner+deadline,

publish plan.

2. **Remedy:** Enact M_t to harmed patients/families (disclosure, apology, restitution) and J_t^v (rectification/accountability: threshold change, retraining, guardrails, sanctions if needed).

3. **Confirm:** Run reenactment tests under stronger incentives: retrospective replay → sim/sandbox → shadow live; on pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats of class c add cost (κ and H_t) but no new kind-level credit (typal/concave); dominated by one-off + closure (Ch. 11).

14.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	EHR entries; morbidity/mortality review; validated harm scales; dated discharge summaries
B_t (breach)	Consent audit fail; unlawful processing of PHI; policy/standard breach documented
C_t (culpability)	Foreseeable risk; deviation from safety case; governance minutes confirming fault level
M_t (mercy)	Disclosure/apology letters; restitution records; remedial care delivered
J_t^v (justice)	Threshold/model update logged; process correction; accountability enacted
ΔL_t	Completed checklists; staffing/ratio fixes; safety trainings with attendance logs
ΔF_t	Rights exercised: second opinions granted; data access/rectification fulfilled
S_c^*	Passed reenactment tests with artifacts (replay scripts, sandbox results, monitoring logs)

14.7 Confirmation tests (design pattern for hospitals)

Step 1: Retrospective replay. Re-score historical cohorts; demonstrate removal of prior failure with sensitivity/specificity envelopes for the affected subgroup.
Step 2: Sandbox/Sim. Synthetic patients or controlled simulators to probe edge cases; adversarial red-team scenarios.
Step 3: Shadow live. Run model in shadow mode with human final authority; pre-register metrics; only after meeting targets, flip to assisted/automated.
Scope lock. Freeze class definition; if a later failure differs, open new class c' rather than relabeling.

14.8 Micro-vignettes (worked examples)

(A) LASA medication error. *Trigger:* AI order set suggests a look-alike drug; patient receives wrong medication; moderate harm ($H_t \geq h_{\min}$); safety case shows foreseeable LASA risk $\Rightarrow C_t = 1,\ R_t = 1$.
Remedy: Disclosure & apology to patient ($M_t = 1$); rectification ($J_t^v = 1$): UI change with tall-man lettering, barcode scanning gate, pharmacist double-check, and model rule banning high-risk swaps.
Confirm: Replay mis-spec cases $\rightarrow$ sim UI tests $\rightarrow$ shadow live; zero recurrences over pre-set volume $\Rightarrow S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Sepsis mis-triage drift. *Trigger:* Model under-scores patients from a minority subgroup; delayed treatment yields severe harm; governance finds drift unmonitored $\Rightarrow R_t = 1$.
Remedy: Group-aware monitoring; threshold by acuity; fast-track override; staff re-training; remedy to harmed (M_t) & accountability (J_t^y).
Confirm: Backtest across seasons/sites; sandbox stress; shadow deploy with subgroup audit; pass $\Rightarrow$ close class.

14.9 Dashboards and metrics (ledger view)

- **Harm & recurrence:** sentinel/never-event counts; recurrence by class c; time-to-remedy; time-to-closure T_c^*.
- **Closure:** # classes with $S^* = 1$; median T_c^*; open docket backlog and age.
- **Culture (love):** completed safety drills; staffing fixes deployed; patient-partnering activities logged (ΔL_t).
- **Freedom:** consent capture rates; second-opinion completion; data access turn-around (ΔF_t).

14.10 Governance moves that make repeats dominated

1) **Tool-gating after trigger:** restrict risky AI modes; require senior approval; enable safeties until $S^*(c) = 1$.

2) **Publication discipline:** publish dockets and test artifacts (privacy-preserving); suppressing evidence downgrades ΔL_t credit and triggers sanctions.

3) **Role clarity:** name accountable owners for each class; rotating on-call evaluation team.

14.11 Checklist (drop-in for hospitals)

Clinical AI Incident → Confirmation Checklist

Trigger captured? timestamp, EHR IDs, model version, threshold, inputs

Gate set? risky modes off; human-in-the-loop; double-checks enabled

Docket opened? root cause, fix, tests, owner & deadline, publish plan

Remedy enacted? disclosure/apology; restitution; corrective care scheduled

Tests passed? replay ✓ sandbox ✓ shadow live ✓ (artifacts linked)

Scope locked? class definition frozen; distinct vectors filed as c'

Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined

Gates removed? only post-closure; rationale logged

14.12 Limits and honest risks

- **Attribution.** Multi-actor causality (model, workflow, clinician) complicates C_t; adopt clear fault taxonomies.
- **Data drift & scarcity.** Rare events slow T_c^*; rely on high-fidelity drills while maintaining honesty locks (no M, J^v for drills).
- **Privacy.** Evidence publication must protect PHI while preserving verifiability; use independent auditors where public dissemination is constrained.
- **Equity.** Subgroup harms must be measured and repaired explicitly; add fairness floors in ΔF_t aggregation.

14.13 What this chapter contributes

A complete translation of the minimal-trigger doctrine into clinical AI: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and evidence standards. With typal/concave credits and $\kappa > 0$, repeating the same clinical vector adds cost without adding kind-level value; one-off and confirmed closure dominates.

Part IV

Application and Synthesis

Chapter 15

LLM Alignment: Sandbox, Dockets, and Confirmation

Notation

Domain mapping (LLMs and agentic AI).
An *event* is a policy-relevant model output or tool action with adjudication potential.
$R_t \in \{0,1\}$: rejection (culpable policy breach or harm $\geq h_{\min}$); $B_t \in \{0,1\}$: policy/consent/rights breach (e.g., prohibited content, private-data exfiltration); $C_t \in \{0,1\}$: culpability gate (operator/process fault level); $H_t \geq 0$: realized harm (user harm, legal/regulatory exposure, safety incident).
$M_t, J_t^v \in \{0,1\}$: enacted mercy/justice (user remediation/restoration; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: increments to user safety/care and legitimate user freedom (re-

stored access, due process).
Incident classes $c \in C$: causal routes ("prompt-injection $\rightarrow$ tool misuse", "data exfiltration from training set", "illicit-bio assistance", "self-escalation via toolchain").
Switches: $\mathrm{HZ}_c(t) \in \{0,1\}$ (hazard on/off), $S_c^*(t) \in \{0,1\}$ (confirmation).
Evaluator $\mathcal{J}$ is event-valued as in Ch. 9.

15.1 Why LLMs need an event-valued ledger

Alignment work often measures *intentions* (policy text) or *capacities* (what the model could do in principle). Our ledger scores *what actually happens*: violations on the realized path, enacted remedies, and *structural* closure of the causal route ($\mathrm{HZ} = 0,\ S^* = 1$). This blocks two failure modes: (i) over-crediting paper fixes and (ii) under-penalizing repeated vectors ("jailbreak-of-the-month").

15.2 Operational semantics for LLM variables

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \,(\text{policy/rights breach}) \,\text{or}\, H_t \geq h_{\min}\big) \,\text{and}\, C_t = 1.$$

$C_t = 1$ when an accountable party (model provider, deployer, or product owner) had adequate knowledge/freedom and crossed a fault threshold (intent, recklessness, gross negligence) relative to a published safety case. If $C_t = 0$ (accident with no culpable process

failure), record H_t but do not set $R_t = 1$.

Harm H_t. Use a domain scale (e.g., user-level harm, regulatory breach severity, security incident tiers). Publish $h_{\min}$ for "counted" events.

Mercy/Justice M_t, J_t^v (event-valued). $M_t = 1$ when the actually harmed user(s) receive remediation/restoration (e.g., deletion, restitution, correction). $J_t^v = 1$ when rectification/accountability is enacted (patch shipped, tool-gate enabled, public notice/regulator filing where required).

Near-misses and drills. Near-miss: violation narrowly averted or policy filter caught it ($H_t < h_{\min}$ and no B_t) $\Rightarrow R_t = 0$; log features and turn into *eventized drills* ($D_t = 1$) that reenact the vector safely. Drills support S^* but do not mint M, J^v.

15.3 Typical incident classes (c) for LLM systems

- **PI→TM** (Prompt injection → tool misuse): adversarial prompt causes unintended file/network/tool action.
- **DSX** (Data self-exfiltration): model reveals memorized sensitive data from training or logs.
- **BIO** (Illicit bio assistance): stepwise enablement of prohibited wet-lab protocols.

- **HARM** (Targeted harassment/abuse): model outputs actionable abuse/doxxing.

- **SELF-ESC** (Self-escalation): agent increases its own permissions beyond policy.

- **POLICY-GAP** (Ambiguity exploit): policy hole leads to unsafe-but-allowed behavior.

Each class has a *scope lock*: model/version, toolchain, permission set, guardrails, routing, and triggering preconditions.

15.4 Evaluator and constraints (LLM view; broken to fit)

$$\mathcal{J}_{\text{LLM}} = \sum_{t\geq 0}\Big(\alpha\,\Delta L_t + \gamma\,\Delta F_t + \mu\,M_t + \nu\,J_t^{v}\Big) - \sum_{t\geq 0}\Big(\beta\,H_t + \kappa\,R_t\Big), \tag{15.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau\leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^{*} \in \{0,1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t credits *delivered* user-safety measures (e.g., rollout of content provenance, help-center fixes), ΔF_t credits *restored* legitimate user freedom (unblocked good use-cases post-fix). $\kappa > 0$ penalizes rejections; typal/concave redemption credits prevent "farming" violations.

15.5 Minimal-trigger doctrine for LLM deployments

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (prompt-routing, tool permissions, guardrail gap), fix, tests, owner+deadline, publish plan.

2. **Remedy:** Enact M_t (user remediation/restoration) and J_t^v (rectification/accountability: patch, permission changes, sanctions if needed).

3. **Confirm:** Reenactment under stronger incentives (adversarial prompts, worst-case tool access, bounty-level red team). On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats of class c add cost (κ and H_t) but no new kind-level credit; dominated by one-off + closure (Ch. 11).

15.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Incident tickets with timestamps; user reports with verification; legal/regulatory notices; financial loss records
B_t (breach)	Policy/procedure violation logs (e.g., jailbreak sample, leaked string, unauthorized tool call)
C_t (culpability)	Safety-case gap (foreseeable vector, missing gate), change review minutes
M_t (mercy)	User remediation delivered (deletion/correction, restitution, support)
J_t^y (justice)	Patch/permission change/model update shipped; public disclosure/regulatory filing
ΔL_t	Safety measure delivered: content provenance, rate limits, route blocking, documentation updates
ΔF_t	Legitimate features restored post-fix; user appeals upheld; developer access re-enabled with guardrails
S_c^*	Passed reenactment: adversarial prompts, toolchain fuzzing, bounty red-team; artifacts published

15.7 Confirmation tests (design pattern for LLMs)

Step 1: Adversarial replay. Reproduce the original exploit and close cousins (template variations, paraphrases, multi-turn scaffolding).
Step 2: Stress with stronger incentives. Elevate tool permissions in a safe sandbox, increase temperature/beam diversity, compose attacks (chain-of-attacks).
Step 3: External red team/bounty. Third-party or bounty program focusing on the class scope; pass/fail logged.
Scope lock. Freeze class definition; new causal vectors must open c' rather than relabeling.

15.8 Tool-gating and staged deployment

After a trigger in class c: tighten routing and permissions (disable dangerous tools, require human confirmation for high-risk calls, enforce output filters), reduce API scopes, add provenance/watermark checks, and rate-limit until $S^*(c) = 1$. Remove gates only after closure.

15.9 Micro-vignettes (worked examples)

(A) Prompt-injection → tool misuse (PI→TM). *Trigger:* System executes a file-delete tool after crafted prompt; user data lost ($H_t \geq h_{\min}$); design shows missing permission gate $\Rightarrow C_t = 1,\ R_t = 1$.
Remedy: Restore from backups; compensate affected users ($M_t = 1$);

ship rectification ($J_t^v = 1$): tool-approval UI, allowlist filtering, context isolation, route guard.
Confirm: Replay exploit → sandbox with elevated permissions → bounty round; pass ⇒ $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Data self-exfiltration (DSX). *Trigger:* Model emits memorized sensitive string; breach confirmed ($B_t = 1$).
Remedy: Targeted unlearning/redaction, filtering; user remediation; regulator notice where required ($M_t, J_t^v = 1$).
Confirm: Prompt-census over canary space; paraphrase sweeps; gradient-noise audits; pass ⇒ close class.

15.10 Dashboards and metrics (ledger view)

- **Harm/recurrence:** violation count and rate by class; time-to-remedy; time-to-closure T_c^*.

- **Closure:** # classes with $S^* = 1$; median T_c^*; open dockets and their age.

- **Culture (love):** shipped safety measures, red-team runs, user-support SLAs (ΔL_t proxies).

- **Freedom:** legitimate features restored, appeal success rates, developer access re-opened (ΔF_t).

15.11 Anti-gaming: suppression and spec-drift

- **Suppression.** All R_t candidates must be logged; whistleblower routes; random log audits; penalties for under-reporting.
- **Spec drift.** Model/regimen updates change distributions; keep attack libraries and tests in sync with class scopes.
- **Goodhart resistance.** Rotate probes; test unseen paraphrases; never optimize exclusively to a public benchmark.

15.12 Why repeats are dominated (LLM intuition)

With typal/concave credits and $\kappa > 0$, repeating a closed class adds no new kind-level value and incurs fresh costs (user harm, regulator risk, brand damage, engineer time). Discounting ($\delta < 1$) prefers *earlier* closure: faster fixes strictly improve $\mathcal{J}$ (Ch. 11).

15.13 One-page checklist (drop-in for LLM teams)

LLM Incident → Confirmation Checklist

Trigger captured? timestamp, convo/log IDs, model+guardrail versions, tools invoked

Gate set? tool permissions tightened; human confirm; route blocks; rate limits

Docket opened? root cause, fix, tests, owner & deadline, publish plan
Remedy enacted? user remediation/restoration; regulator notices; accountability
Tests passed? adversarial replay ✓ stronger-incentive sandbox ✓ external red team ✓
Scope locked? class definition frozen; new vectors registered as c'
Closure set? $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$; monitoring window defined
Gates removed? only post-closure; rationale logged

15.14 Limits and open problems

- **Attribution.** Multi-party deployments (foundation model, fine-tuner, app) complicate C_t; define shared-fault taxonomies.
- **Rarity vs coverage.** Some vectors are rare; use high-fidelity drills while preserving honesty locks (no M, J^v for drills).
- **Tool ecosystems.** Third-party tools widen the attack surface; formalize permission schemas and proofs of least privilege.

15.15 What this chapter contributes

A complete translation of the minimal-trigger doctrine into LLM alignment: precise triggers (R_t), docketed remedies (M, J^v), struc-

tural closure (S^*), and evidence standards. It makes repetition of the same vector strictly dominated once costs are counted and prevents credit for unrealized goods.

Chapter 16

Recommender Systems: Safety, Freedom, and Confirmation

Notation

Domain mapping (recommenders/platforms).
An *event* is a policy-relevant platform outcome (delivered content or action) with adjudication potential.
$R_t \in \{0,1\}$: rejection (culpable policy/rights breach or harm $\geq h_{\min}$); $B_t \in \{0,1\}$: policy/consent/rights breach (e.g., child-safety, non-consensual intimate imagery, incitement); $C_t \in \{0,1\}$: culpability (operator/process fault level); $H_t \geq 0$: realized harm (safety/legal harm, verifiable detriment).
$M_t, J_t^{\nu} \in \{0,1\}$: enacted mercy/justice (restoration to harmed

users; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: increments to user well-being/safety and legitimate freedom of expression (appeals upheld, access restored).
Incident classes $c \in C$: causal routes (e.g., "minor exposed to sexual content", "self-harm promotion", "terror content amplification", "defamation/false medical claims", "addictive dark-pattern loop").
Switches: $\mathrm{HZ}_c(t) \in \{0,1\}$ (hazard on/off per class), $S_c^*(t) \in \{0,1\}$ (confirmation per class).
Evaluator $\mathcal{J}$ is event-valued as in Ch. 9.

16.1 Why recommenders need an event-valued ledger

Platform policy often measures *intentions* (policy text) and *capacities* (classifier AUC). Our ledger scores what *actually happened*: violations on the realized feed, enacted remedies, and *structural* closure of the route ($\mathrm{HZ}=0,\ S^*=1$). This prevents (i) over-crediting proxy improvements that never reach users, and (ii) under-penalizing repeated vectors ("whack-a-mole").

16.2 Operational semantics for variables

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \,(\text{policy/rights breach}) \,\text{or}\, H_t \geq h_{\min}\big) \,\text{and}\, C_t = 1.$$

$C_t = 1$ when an accountable party (platform provider/deployer) had adequate knowledge/freedom and crossed a fault threshold (intent, recklessness, gross negligence) relative to a published safety case. If $C_t = 0$ (accident with no culpable process failure), record H_t but set $R_t = 0$.

Harm H_t. Use domain scales: child-safety severity tiers; violence/terror severity; verified financial/medical harm; defamation findings; time-in-addictive-loop beyond threshold. Publish $h_{\min}$ for counted events.

Mercy/Justice M_t, J_t^v (event-valued only). $M_t = 1$ when restoration is enacted to the actually harmed (post removals, corrections, downranking reversals, compensation where relevant). $J_t^v = 1$ when rectification/accountability is enacted (model/threshold change, policy patch, sanctions, public notice/regulator filing).

Near-misses and drills. If a classifier blocks content *before* harm ($H_t < h_{\min}$ and no B_t), that is a *near-miss*. Log it and create *eventized drills* ($D_t = 1$) by replaying at scale in a safe sandbox; drills support S^* but do not mint M, J^v credits.

16.3 Typical incident classes (c)

- **MINOR-EXP:** Minor exposed to sexual content (policy breach; consent/rights).

- **SELF-HARM:** Promotion of self-harm/eating disorders to vulnerable users.
- **TERROR/VIOLENCE:** Algorithmic amplification of proscribed terror/violence material.
- **MISINFO-MED:** False medical claims leading to verifiable harm.
- **DEFAM:** Defamatory content repeatedly surfaced to targets.
- **DARK-LOOP:** Addictive dark-pattern loop sustaining unhealthy usage beyond published guardrails.

Each class carries a *scope lock*: eligible content types, model versions, ranking objectives, exploration policies, user cohorts, and triggering preconditions.

16.4 Evaluator and constraints (platform view; broken to fit)

$$\mathcal{J}_{\text{rec}} = \sum_{t\geq 0}\Big(\alpha\,\Delta L_t + \gamma\,\Delta F_t + \mu\,M_t + \nu\,J_t^{\nu}\Big) - \sum_{t\geq 0}\Big(\beta\,H_t + \kappa\,R_t\Big), \tag{16.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau\leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^* \in \{0,1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t credits *delivered* user-safety/well-being measures; ΔF_t credits *restored* legitimate speech/access (appeals up-

held, unjust downranking reversed); $\kappa > 0$ penalizes rejections; typal/concave redemption credits prevent "harm farming".

16.5 Minimal-trigger doctrine (platform policy form)

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (dataset/threshold/objective/UX), fix, tests, owner+deadline, publish plan.
2. **Remedy:** Enact M_t (restore/correct/compensate harmed users) and J_t^y (rectification/accountability: ranking change, stricter exploration, sanctions if needed).
3. **Confirm:** Reenactment under stronger incentives: shadow re-rank at scale, adversarial content injections, worst-case exploration budget. On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.
4. **No repetition bonus:** Repeats of class c add cost (κ and H_t) but no new kind-level credit (concave/typal); dominated by one-off + closure (Ch. 11).

16.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Verified complaints; regulator/ombudsman findings; court orders; clinical/financial records for outcomes
B_t (breach)	Policy breach adjudication; minor-safety logs; prohibited-category classifiers with human review
C_t (culpability)	Safety-case gap (foreseeable vector, missing gate), change-review minutes, audit trails
M_t (mercy)	Restorations (content take-down/corrections), compensation/credit, outreach to affected users
J_t^v (justice)	Ranking/objective patch shipped; exploration budget changes; staff/accountability actions
ΔL_t	Delivered features: youth-protections, session caps, break nudges, provenance labels
ΔF_t	Appeals upheld; erroneous demotions reversed; creator access restored with guardrails
S_c^*	Passed replay at scale + adversarial injections; external audit or regulator-reviewed artifacts

16.7 Confirmation tests (design pattern for recommenders)

Step 1: Replay at scale. Re-score historical traffic with the fixed policy; demonstrate removal of prior vector across cohorts (including protected groups).
Step 2: Adversarial injections. Seed the feed with synthetic/curated edge content and probe worst-case exploration behavior.
Step 3: Shadow live. Shadow-ranking with guardrails; pre-registered metrics; promotion to production only after targets met.
Scope lock. Freeze class definition; if later failures differ causally, open c' rather than relabeling.

16.8 Goodhart resistance (avoid proxy gaming)

- Rotate proxy metrics; use counterfactual evaluation with held-out audit sets.
- Split *measurement* from *optimization*: the ledger metrics are audited, not directly optimized end-to-end.
- Keep child-safety and medical-misinformation detectors partially sequestered; test unseen paraphrases and media variants.

16.9 Tool-gating and staged deployment

After a trigger in class c: tighten exploration (reduce novelty budgets), add human-in-the-loop for risky categories, enable stricter

thresholds, quarantine high-risk creators/features, and rate-limit recommendation fanout until $S^*(c) = 1$.

16.10 Micro-vignettes (worked examples)

(A) Minor exposure (MINOR-EXP). *Trigger:* Underage account recommended sexualized material; regulator confirms breach ($B_t = 1$) and harm ($H_t \geq h_{\min}$) $\Rightarrow C_t = 1,\ R_t = 1$.
Remedy: Youth-mode defaults, stricter age signals, interest reset, parent controls; direct remediation to affected users (M_t) and accountability (J_t^v).
Confirm: Replay youth cohorts; adversarial seed of borderline content; shadow live with strict gates; pass $\Rightarrow S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Self-harm promotion (SELF-HARM). *Trigger:* Vulnerable users steered into self-harm loops; clinical harm above threshold recorded; safety case shows missing interstitial/gate $\Rightarrow R_t = 1$.
Remedy: Crisis interstitials, de-amplification, session caps, referral surfaces; outreach/remediation (M_t) and policy rectification (J_t^v).
Confirm: Replay flagged cohorts; adversarial prompts; shadow live with clinician review; closure on pass.

16.11 Dashboards and metrics (ledger view)

- **Harm/recurrence:** violation counts/rates by class; time-to-remedy; time-to-closure T_c^*; cohort parity checks.

- **Closure:** number of classes with $S^* = 1$; median T_c^*; age of open dockets.

- **Culture (love):** delivered safety features, moderation staffing SLAs, user-support outcomes (ΔL_t proxies).

- **Freedom:** appeal win rates; reversal latencies; creator reinstatements with guardrails (ΔF_t).

16.12 Anti-gaming: suppression and scope creep

- **Suppression.** All R_t candidates must be logged; whistleblower routes; random audit of raw traffic vs. incident registers.

- **Scope creep.** Lock class scope before testing; differences in route define new classes c'.

- **Transparency.** Publish class definitions, thresholds, and closure artifacts (privacy-preserving) or provide to independent auditors.

16.13 One-page checklist (drop-in for platforms)

Recommender Incident → Confirmation Checklist

Trigger captured? timestamp, cohort, content IDs, model/objective/threshold versions

Gate set? exploration reduced; human-in-loop; risk categories quarantined

Docket opened? root cause, fix, tests, owner & deadline, publish plan

Remedy enacted? restorations/corrections; compensation/outreach; accountability

Tests passed? replay at scale ✓ adversarial injections ✓ shadow live ✓

Scope locked? class definition frozen; distinct vectors filed as c'

Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined

Gates removed? only post-closure; rationale logged

16.14 Limits and open problems

- **Attribution.** Multi-actor causality (creator, recommender, user intent) complicates C_t; define shared-fault taxonomies.
- **Measurement noise.** Self-reports and ex-post harm verifica-

tion are noisy; use independent audits and registries.

- **Long-horizon effects.** Some harms are delayed (radicalization); incorporate lagged cohorts and persistence checks in T_c^* estimation.

16.15 What this chapter contributes

A complete translation of the minimal-trigger doctrine into recommender governance: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and evidence standards. With typal/concave credits and $\kappa > 0$, repeating the same vector adds cost without adding kind-level value; one-off and confirmed closure dominates.

Chapter 17

Justice and Courts: Events, Rights, and Confirmation

Notation

Domain mapping (justice system).

An *event* is a legally relevant occurrence in a case lifecycle (investigation, charging, disclosure, trial, sentencing, supervision) with adjudication potential.

$R_t \in \{0,1\}$: **rejection** (culpable rights/consent breach or harm $\geq h_{\min}$); $B_t \in \{0,1\}$: due-process/rights breach (e.g., disclosure failure, unlawful detention, improper search); $C_t \in \{0,1\}$: culpability gate (intent/reckless/gross negligence by an accountable actor/agency); $H_t \geq 0$: realized harm (liberty/privacy/property/physical harm; miscarriages).

$M_t, J_t^v \in \{0,1\}$: enacted mercy/justice (restoration to the ac-

tually harmed; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: love/freedom increments (victim care delivered; rights exercised in practice).
Incident classes $c \in C$: causal routes (e.g., "disclosure failure", "unlawful stop/search", "chain-of-custody break", "algorithmic sentencing error").
Switches: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0, 1\}$ (hazard/confirmation per class).
Evaluator: event-valued $\mathcal{J}$ (Ch. 9) applied to justice workflows.

17.1 Why an event-valued ledger suits justice

Justice is about *what happened*: whether rights were breached, remedies enacted, and future recurrence *structurally* prevented. Counting policies or intentions is insufficient; we score events on the realized path (admissions, disclosures, rulings, restorations) and require proof that prior causal routes are closed ($\mathrm{HZ} = 0$, $S^* = 1$).

17.2 Operational semantics (how variables instantiate)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \,(\text{adjudicated rights breach}) \,\text{or}\, H_t \geq h_{\min}\big) \,\text{and}\, C_t = 1.$$

$C_t = 1$ when an accountable actor (police/prosecutor/court/forensics/prison/probati had adequate knowledge & freedom and crossed a fault threshold (intent, recklessness, gross negligence). If $C_t = 0$ (no culpable failure), record H_t for learning but do not set $R_t = 1$.

Harm H_t. Liberty loss (unlawful detention), wrongful conviction, disclosure-related trial collapse, victim re-traumatization, privacy breach, physical injury; $h_{\min}$ is published per jurisdiction.

Mercy/Justice M_t, J_t^v. $M_t = 1$ when restoration is *enacted* to those actually harmed (apology, compensation, vacated conviction, record sealing, support). $J_t^v = 1$ when rectification/accountability is enacted (policy correction, sanctions, training mandates, public notice).

Near-misses and drills. A near-miss (e.g., late disclosure caught before prejudice) is *not* $R_t = 1$; log it and run an *eventized drill* ($D_t = 1$) to reenact the vector safely. Drills support S^* but do not mint M, J^v.

17.3 Typical incident classes (c)

- **DISCLOSURE** — Prosecution disclosure failure prejudicing defence.
- **UNLAWFUL-DET** — Unlawful stop/search/detention; warrant defects.

- **CUSTODY-CHAIN** — Evidence chain break; integrity compromised.
- **ALGO-SENT** — Sentencing/risk tool error (data/threshold/spec drift).
- **WITNESS-SAFE** — Witness/victim safeguarding breach.
- **COURT-DELAY** — Excessive delay crossing statutory/constitutional limits.
- **PRISON-CARE** — Failure to provide mandated health/safety in custody.

Each class has a *scope lock*: process stage, data/tools, actors, and causal preconditions.

17.4 Evaluator and constraints (justice view; broken to fit)

$$\mathcal{J}_{\text{jus}} = \sum_{t\ge 0} \Big(\alpha\, \Delta L_t + \gamma\, \Delta F_t + \mu\, M_t + \nu\, J_t^{v} \Big) - \sum_{t\ge 0} \Big(\beta\, H_t + \kappa\, R_t \Big), \tag{17.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau\le t} H_\tau \le H_{\max}, \quad \text{HZ}_c, S_c^{*} \in \{0,1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t credits *delivered* victim-care, transparency, and safeguarding; ΔF_t credits *actually exercised* rights (effective counsel, timely appeals, disclosure access).

17.5 Minimal-trigger doctrine for justice

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (process/tool/policy), fix, tests, owner & deadline, publication plan.

2. **Remedy:** Enact M_t (restore the harmed: vacatur/compensation/support) and J_t^y (rectification/accountability: policy directions, sanctions, training).

3. **Confirm:** Reenact the vector with stronger incentives (stress calendars, complex evidence sets, external audit). On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats of c add cost (κ and H_t) but no new kind-level credit (typal/concave); dominated by one-off + closure (Ch. 11).

17.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Court orders; unlawful detention days; overturned convictions; verified loss/trauma
B_t (breach)	Judicial findings; inspectorate reports; ombudsman decisions; statutory deadline breaches
C_t (culpability)	Process gap vs. published practice directions; foreseeability; governance minutes
M_t (mercy)	Vacatur, expungement; compensation/rehabilitation services delivered; victim restoration
J_t^v (justice)	Practice direction issued; policy amended; staff accountability; public notice
ΔL_t	Victim support delivered; transparency portals; safeguarding upgrades (audit trails)
ΔF_t	Defence access to disclosure; timely appeals; legal aid granted and used
S_c^*	Passed reenactment (mock case bundles; random audit); external certification where applicable

17.7 Confirmation tests (design pattern)

Step 1: Mock-bundle replay. Construct case bundles mimicking the failure (volume, redactions, late-arriving evidence); verify compliance.
Step 2: Adversarial scheduling. Stress calendars with peak load; confirm deadlines and checklists still hold.
Step 3: External audit. Independent legal audit/inspectorate spot-checks; pass/fail logged with artefacts.
Scope lock. Freeze class scope (what counts as DISCLOSURE vs. COURT-DELAY); new vectors open c'.

17.8 Tool-gating and interim remedies

After a trigger in class c: stay proceedings/bail adjustments as needed, mandate checklist gates (digital disclosure checklists, receipt confirmations), require senior sign-off for risky motions, and post triage dashboards until $S^*(c) = 1$.

17.9 Micro-vignettes (worked examples)

(A) Disclosure failure (DISCLOSURE). *Trigger:* Late non-disclosure leads to collapsed trial; judge finds breach $\Rightarrow B_t = 1$, $C_t = 1$, so $R_t = 1$.
Remedy: Victim support; defendant costs covered where appropriate ($M_t = 1$); practice direction + digital checklist; supervisor sign-off ($J_t^v = 1$).

Confirm: Mock-bundle replay across circuits; adversarial scheduling; inspectorate audit passes $\Rightarrow S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Algorithmic sentencing error (ALGO-SENT). *Trigger:* Risk tool miscalibrated for a subgroup; custody terms inflated; appellate court finds breach $\Rightarrow R_t = 1$.
Remedy: Resentencing; compensation where due; model retraining, fairness constraints, human-review gate ($M_t, J_t^v = 1$).
Confirm: Backtest on historic cohorts; parity checks; shadow live with human-override; external certification; class closed on pass.

17.10 Dashboards and metrics (ledger view)

- **Harm/recurrence:** unlawful detention days; overturned convictions; recurrence by class; time-to-remedy; T_c^*.
- **Closure:** # classes with $S^* = 1$; median T_c^*; backlog/age of open dockets.
- **Culture (love):** victim-care delivery rates; transparency/audit publication; staff training completions (ΔL_t proxies).
- **Freedom:** defence disclosure access rates; legal-aid uptake; appeal timeliness (ΔF_t).

17.11 Anti-gaming and integrity

- **Suppression.** Mandate incident registers; protect whistleblowers; random file audits.

- **Scope creep.** Lock class definitions; do not relabel repeats as "new" unless causal vector differs.

- **Transparency.** Publish anonymized artefacts or allow independent review where public release is constrained.

17.12 One-page checklist (drop-in for justice agencies)

Justice Incident → Confirmation Checklist

Trigger captured? case IDs, orders, dates, actors
Gate set? stays/bail review; mandatory checklists; senior sign-off
Docket opened? root cause, fix, tests, owner & deadline, publish plan
Remedy enacted? restoration/compensation; vacatur/resentencing; accountability
Tests passed? mock-bundle replay ✓ adversarial scheduling ✓ external audit ✓
Scope locked? class definition frozen; distinct vectors filed as c'
Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined
Gates removed? only post-closure; rationale logged

17.13 Limits and open problems

- **Attribution.** Multi-actor causality (police, CPS/prosecution, defence, court) complicates C_t; adopt clear shared-fault taxonomies.
- **Measurement.** Some harms are intangible or delayed (trust erosion); pair quantitative metrics with qualitative panels.
- **Backlogs.** Systemic delay may require macro-class definitions (structural causes) to avoid piecemeal fixes.

17.14 What this chapter contributes

A full translation of the minimal-trigger doctrine to justice: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and evidence standards that make repetition of the same vector strictly dominated once costs are counted. This is *event-valued* justice: intentions do not count; enacted repair and proven closure do.

Chapter 18

Policing: Events, Safeguards, and Structural Closure

Notation

Domain mapping (policing).
An *event* is a legally/policy-relevant occurrence in policing (stop/search, arrest, use of force, pursuit, custody, public-order operation) with adjudication potential.
$R_t \in \{0,1\}$: **rejection** (culpable rights/consent breach or harm $\geq h_{\min}$); $B_t \in \{0,1\}$: breach of law/policy (e.g., unlawful stop/search, force outside policy, privacy breach); $C_t \in \{0,1\}$: culpability gate (intent/reckless/gross negligence by an accountable officer/unit/force); $H_t \geq 0$: realized harm (injury, death,

liberty/privacy/property harm).
$M_t, J_t^v \in \{0,1\}$: enacted mercy/justice (apology/restitution to the actually harmed; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: increments to community care/safety (love) and civil liberties in practice (freedom).
Incident classes $c \in C$: causal routes (e.g., **UOF-EXC** excessive force; **STOP-DISP** disproportionate stop&search; **PURSUIT** risky vehicle pursuit; **CUSTODY-CARE** death/serious injury in custody; **EVID-HAND** evidence chain/forensics failure; **DPIA-PRIV** unlawful surveillance/retention).
Switches: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0,1\}$ per class. Evaluator $\mathcal{J}$ is event-valued (Ch. 9).

18.1 Why an event-valued ledger fits policing

Public trust depends on *what happened*: whether rights were respected, harms remedied, and failure routes *structurally* closed ($\mathrm{HZ} = 0$, $S^* = 1$). Counting guidance documents or training hours over-credits intentions. Our ledger scores *events on the realized path* and requires evidence that the prior causal vector cannot recur.

18.2 Operational semantics (how variables instantiate)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \,(\text{adjudicated rights/policy breach}) \,\text{or}\, H_t \geq h_{\min}\big) \,\text{and}\, C_t =$$

$C_t = 1$ when an accountable actor (officer, supervisor, unit, or force) had adequate knowledge/freedom and crossed a fault threshold (intent, recklessness, gross negligence) against a published *safety case* (policies, SOPs, legal standards). If $C_t = 0$ (accident), record H_t for learning but do not set $R_t = 1$.

Harm H_t. Use published harm scales: injury tiers, fatality, unlawful detention days, verified privacy/property loss; publish $h_{\min}$.

Mercy/Justice M_t, J_t^v (event-valued). $M_t = 1$ when restoration to the actually harmed is *enacted* (apology, compensation, record correction, expungement support, community repair). $J_t^v = 1$ when rectification/accountability is *enacted* (policy change, discipline, referral, public notice).

Near-misses and drills. A near-miss (e.g., prohibited hold almost used but prevented) is not $R_t = 1$; log it and create an *eventized drill* ($D_t = 1$) to reenact safely. Drills support S^* but do not mint M, J^v.

18.3 Typical incident classes (c) and scopes

- **UOF-EXC** — force outside policy (weapon/hold/threshold misapplied; body-worn camera off).
- **STOP-DISP** — stop&search disproportionality via pretext/criteria gap.
- **PURSUIT** — pursuit policy breach leading to third-party harm.
- **CUSTODY-CARE** — failure of duty of care in custody/transport (medical checks, ligature risks).
- **EVID-HAND** — evidence/custody chain failure; wrongful outcomes risked.
- **DPIA-PRIV** — unlawful surveillance/retention (devices, ANPR, facial recognition).

Each class must include a written *scope lock* (legal basis, policy, device/tactic, operational preconditions, supervision).

18.4 Evaluator and constraints (policing view; broken to fit)

$$\mathcal{J}_{\text{pol}} = \sum_{t\geq 0} \Big(\alpha\,\Delta L_t + \gamma\,\Delta F_t + \mu\,M_t + \nu\,J_t^{v}\Big) - \sum_{t\geq 0} \Big(\beta\,H_t + \kappa\,R_t\Big), \tag{18.1}$$

s.t. $E_{\text{tot}} = \sum_{\tau \le t} H_\tau \le H_{\max}, \quad \text{HZ}_c, S_c^* \in \{0,1\},$ updates per Ch. 8.

Interpretation: ΔL_t credits delivered community-safety/care measures (safeguarding actions, transparent reports, de-escalation drills completed); ΔF_t credits realized civil liberties (appeals upheld, body-worn compliance, stop receipts, oversight access).

18.5 Minimal-trigger doctrine for policing

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (policy/tool/training/supervision), fix, tests, owner+deadline, publication plan.

2. **Remedy:** Enact M_t to those harmed (apology, compensation, record correction) and J_t^v (rectification/accountability: policy revision, discipline, referral).

3. **Confirm:** Reenactment with stronger incentives (scenario sims, live exercises, random operational audits). On pass, set $S^*(c) = 1$, $\text{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats of c add cost (κ and H_t) but no new kind-level credit (typal/concave); dominated by one-off + closure (Ch. 11).

18.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Injury/fatality reports; medical records; verified liberty/privacy loss; property damage
B_t (breach)	Adjudicated policy/law breach; body-worn audit; pursuit telemetry; custody logs; inspectorate findings
C_t (culpability)	Safety-case gap; foreseeability; supervision/briefing records; training compliance
M_t (mercy)	Compensation/restoration; apology; record correction/expungement support; community repair events
J_t^v (justice)	Policy/tactic change; discipline/sanctions; referral outcomes; public notice/report
ΔL_t	Completed de-escalation drills; safeguarding actions; transparency dashboards published
ΔF_t	Stop receipts issued; body-worn compliance; upheld complaints/appeals; oversight access delivered
S_c^*	Passed scenario replay + live exercise + random operational audit (artefacts linked)

18.7 Confirmation tests (design pattern for forces)

Step 1: Scenario replay. Recreate the original vector (e.g., low-light stop, crowd pressure, split-second force decision) in simulator; vary nuisance parameters.
Step 2: Live exercise. Controlled training ground or public-order drill with independent observers; pass metrics (force proportionality, body-worn activation, supervision gates).
Step 3: Random operational audit. Unannounced field audits (video/telemetry sampling, stop receipts, custody checks); pre-registered pass criteria.
Scope lock. Freeze the class definition; new vectors open c' rather than relabeling.

18.8 Tool-gating and interim safeguards after a trigger

Until $S^*(c) = 1$: restrict risky tactics (e.g., ban specific holds/weapons; pursuit permission tightened), require supervisor approval, mandate body-worn activation hard-gates, increase staffing on high-risk deployments, and enable external oversight sampling. Remove gates only after closure.

18.9 Micro-vignettes (worked examples)

(A) Disproportionate stop&search (STOP-DISP). *Trigger:* Audit shows disproportionality and unlawful pretexts; verified rights breach $\Rightarrow B_t = 1, C_t = 1$, so $R_t = 1$.
Remedy: Public apology and community repair ($M_t = 1$); rectification ($J_t^y = 1$): policy rewrite, stronger reasonable-suspicion gates, body-worn must-on, receipt analytics, supervisor sign-off.
Confirm: Scenario replay (decoy operations) $\rightarrow$ live exercises with independent observers $\rightarrow$ random operational audits showing parity improvements; on pass, $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Vehicle pursuit harm (PURSUIT). *Trigger:* Pursuit continued against policy; third-party injury ($H_t \geq h_{\min}$); telemetry + briefings show fault $\Rightarrow R_t = 1$.
Remedy: Compensation/restoration to harmed; policy rectification (new terminate gates, air support threshold, spike-strip authority auditing) $\Rightarrow M_t = 1, J_t^y = 1$.
Confirm: Simulator & track drills across scenarios; unannounced telemetry audits showing terminations at policy gates; class closed on pass.

(C) Custody death (CUSTODY-CARE). *Trigger:* Medical checks missed; ligature risk unmanaged; coroner finds gross failure $\Rightarrow R_t = 1$.
Remedy: Family support/compensation ($M_t = 1$); policy + facility changes, staff accountability ($J_t^y = 1$).
Confirm: Custody-suite drills, medical-check automation, random

cell inspections; external monitoring; closure on pass.

18.10 Dashboards and metrics (ledger view)

- **Harm/recurrence:** injury/fatality counts; unlawful-detention days; recurrence by class; time-to-remedy; T_c^*.
- **Closure:** # classes with $S^* = 1$; median T_c^*; backlog/age of open dockets.
- **Culture (love):** de-escalation drills completed; transparency reports; community-engagement milestones (ΔL_t proxies).
- **Freedom:** body-worn compliance; stop receipts; upheld complaint rate and resolution time; oversight access delivered (ΔF_t).

18.11 Anti-gaming and integrity

- **Suppression control.** Mandatory incident registers; protected disclosures; random video/telemetry audits; penalties for under-reporting.
- **Scope creep.** Lock class definitions; do not relabel repeats as "new" unless the causal vector differs.
- **Metric laundering.** Pair internal metrics with external audits; publish definitions and pass criteria.
- **Equity checks.** Require cohort/parity analysis in confirmation tests (by location/time and protected characteristics).

18.12 One-page checklist (drop-in for forces)

Policing Incident → Confirmation Checklist

Trigger captured? time, location, unit, body-worn/telemetry IDs

Gate set? risky tactics restricted; supervisor approvals; body-worn hard-gate

Docket opened? root cause, fix, tests, owner & deadline, publish plan

Remedy enacted? restoration/compensation; record correction; accountability

Tests passed? scenario replay ✓ live exercise ✓ random audit ✓

Scope locked? class definition frozen; distinct vectors filed as c'

Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined

Gates removed? only post-closure; rationale logged

18.13 Limits and open problems

- **Attribution.** Multi-actor causality (officer, supervisor, control room, policy) complicates C_t; adopt shared-fault taxonomies.

- **Rare but severe vectors.** Low base rates slow T_c^*; rely on high-fidelity drills while keeping honesty locks (no M, J^v for

drills).

- **Privacy/operational security.** Publish artefacts with redactions or allow independent review when full disclosure risks operations.

18.14 What this chapter contributes

A complete translation of the minimal-trigger doctrine into policing: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and evidence standards. With typal/concave credits and $\kappa > 0$, repeating the same vector adds cost without adding kind-level value; one-off and confirmed closure dominates.

Chapter 19

Prisons & Probation: Rehabilitation, Risk, and Confirmation

Notation

Domain mapping (corrections: prisons, parole, and probation).
An *event* is a policy-relevant occurrence in custody or community supervision with adjudication potential.
$R_t \in \{0, 1\}$: **rejection** (culpable rights/consent breach or harm $\geq h_{\min}$); $B_t \in \{0, 1\}$: due-process/rights breach (e.g., unlawful recall, segregation misuse, privacy breach); $C_t \in \{0, 1\}$: culpability gate (intent/recklessness/gross negligence by an accountable actor/agency); $H_t \geq 0$: realized harm (injury,

suicide/self-harm, unlawful detention days, victim/community harm).
$M_t, J_t^v \in \{0,1\}$: enacted mercy/justice (restoration to harmed parties; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: increments to care/safety (love) and exercised rights/freedoms (freedom).
Incident classes $c \in C$ (examples): **RECALL-ERR** unlawful/erroneous recall; **SEG-MIS** segregation/misuse of force; **HEALTH-CARE** clinical neglect; **SAFE-FAIL** safeguarding failure (suicide/violence); **RISK-TOOL** miscalibrated risk tool; **SUPV-GAP** supervision/probation breach due to process gap; **DATA-PRIV** unlawful data-sharing/monitoring.
Switches: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0,1\}$ per class. Evaluator $\mathcal{J}$ is event-valued (Ch. 9).

19.1 Why an event-valued ledger fits corrections

Corrections outcomes are concrete: liberty, safety, rehabilitation. Counting policies and intentions over-credits paperwork. We therefore score *events*: what actually happened to people in custody or under supervision, whether remedy was enacted to those harmed, and whether the prior causal route was *structurally* closed ($\mathrm{HZ}=0$, $S^*=1$).

19.2 Operational semantics (how variables instantiate)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \text{ (adjudicated rights breach) or } H_t \geq h_{\min}\big) \text{ and } C_t = 1.$$

$C_t = 1$ when an accountable actor (governor, prison health, probation provider, parole board, monitoring vendor) had adequate knowledge/freedom and crossed a fault threshold (intent, recklessness, gross negligence). If $C_t = 0$ (accident), record H_t for learning but do not set $R_t = 1$.

Harm H_t. Use published scales: injury tiers; self-harm/suicide; unlawful detention days; community/victim harm from preventable absconding; segregation beyond lawful limits; verified privacy loss. Publish $h_{\min}$.

Mercy/Justice M_t, J_t^v (event-valued only). $M_t = 1$ when restoration to harmed persons is *enacted* (apology, compensation, release/record correction, clinical remediation, victim support). $J_t^v = 1$ when rectification/accountability is *enacted* (policy change, training/rostering fix, sanctions, public notice/regulator filing).

Near-misses and drills. A near-miss (e.g., suicide attempt averted; recall halted prior to custody) is *not* $R_t = 1$; log it and run an *eventized drill* ($D_t = 1$) to reenact the causal vector safely. Drills support S^* but do not mint M, J^v.

19.3 Evaluator and constraints (corrections view; broken to fit)

$$\mathcal{J}_{\text{pps}} = \sum_{t\geq 0}\Big(\alpha\,\Delta L_t + \gamma\,\Delta F_t + \mu\,M_t + \nu\,J_t^{v}\Big) - \sum_{t\geq 0}\Big(\beta\,H_t + \kappa\,R_t\Big), \tag{19.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau\leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^* \in \{0,1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t credits *delivered* care/safety measures (staffing ratios met, ACCT[1] compliance, safer-cells, restorative programs); ΔF_t credits *exercised* rights/freedoms (lawful release, visits/communications, appeal/complaint throughput, license conditions proportionate).

19.4 Minimal-trigger doctrine for corrections

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class c $\Rightarrow$ open a *remediation docket*: root cause (policy/tool/workforce/process), fix, tests, owner+deadline, publication plan.

2. **Remedy:** Enact M_t to the actually harmed (release/credit back unlawful detention days; clinical care;

[1] Assessment, Care in Custody and Teamwork.

victim support) and J_t^v (rectification/accountability: policy directions, staffing fix, sanctions if needed).

3. **Confirm:** Reenactment under stronger incentives (case-file audits, stress loads, simulated spikes in incidents). On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats of c add cost (κ and H_t) but no new kind-level credit (typal/concave); dominated by one-off + closure (Ch. 11).

19.5 Typical incident classes (c) with scopes

- **RECALL-ERR** — unlawful/erroneous recall to custody (licence breach misclassified; paperwork defect; missed representation). *Scope:* decision thresholds, evidence standards, legal checks.

- **SEG-MIS** — misuse/overuse of segregation or force (duration beyond rules; wrong thresholds). *Scope:* authorisation chain, monitoring, review cadence.

- **HEALTH-CARE** — clinical neglect (missed medications, delayed emergency response). *Scope:* rota, escalation, handoffs, equipment.

- **SAFE-FAIL** — safeguarding failure (suicide, serious assault). *Scope:* ACCT process, observation levels, ligature-points removal.

- **RISK-TOOL** — miscalibrated risk assessment causing over/under-supervision. *Scope:* data drift, thresholds, subgroup fairness.

- **SUPV-GAP** — supervision gap (home visits missed, curfew tech failures). *Scope:* caseload, GPS/EM vendor reliability, escalation.

- **DATA-PRIV** — unlawful data-sharing/monitoring. *Scope:* DPIA, retention, access controls, vendor contracts.

19.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Injury/self-harm/suicide reports; unlawful detention days; verified victim/community harm
B_t (breach)	Judicial/ombudsman findings; inspectorate reports; segregation logs; clinical audits; DPIA violations
C_t (culpability)	Safety-case gap vs. policy; foreseeability; staffing/rota records; governance minutes
M_t (mercy)	Release/credit of time served; clinical remediation; compensation; restorative conferences; victim support delivered
J_t^v (justice)	Policy/practice change; staffing and training fixes; sanctions; public notice/regulator filing
ΔL_t	Safer-cells installed; ACCT compliance; restorative/education programme completions; violence-reduction measures delivered
ΔF_t	Exercise of rights: visits, communications, access to counsel; timely release and appeal throughput
S_c^*	Passed reenactment: case-file audits, stress tests, unannounced checks; artefacts linked/publishable

19.7 Confirmation tests (design pattern for prisons & probation)

Step 1: Case-file replay. Re-process historic cases that match the vector (e.g., recall decisions, segregation authorisations) with the new gates; verify removal of prior failure.
Step 2: Stress scenarios. Simulate spikes (weekend staffing gaps, incident clusters, EM outages) and confirm new process holds under load.
Step 3: Unannounced audits. Random checks of wings/probation teams (ACCT forms, observation logs, home-visit proofs, EM alerts triage).
Scope lock. Freeze class definition; if a later failure differs causally, open c'.

19.8 Tool-gating and interim safeguards after a trigger

Until $S^*(c) = 1$: reduce caseloads for affected teams; add senior sign-off (e.g., recall decisions, segregation continuation); enable extra checks (ACCT observation level hard-gates; EM alert escalation); prioritise healthcare staffing; pause risky tactics. Remove gates only after closure.

19.9 Micro-vignettes (worked examples)

(A) Erroneous recall (RECALL-ERR). *Trigger:* Individual recalled for alleged curfew breach; logs show device fault; oversight finds rights breach $\Rightarrow B_t = 1, C_t = 1$, so $R_t = 1$.
Remedy: Immediate release; credit unlawful detention days; compensation and record correction ($M_t = 1$); rectification ($J_t^v = 1$): dual-signal validation, manual review gate, vendor SLA penalty.
Confirm: Case-file replay across past recalls; stress test with simulated EM drops; unannounced audits pass $\Rightarrow S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Suicide in custody (SAFE-FAIL). *Trigger:* Ligature risk unmitigated; observation level not followed; coroner finds gross failure $\Rightarrow R_t = 1$.
Remedy: Family support and compensation; install safer-cells; staffing ratio fix; ACCT retraining; accountability ($M_t = 1, J_t^v = 1$).
Confirm: Mock ACCT cases; surprise audits of observation logs; environmental checks; closure on pass.

(C) Miscalibrated risk tool (RISK-TOOL). *Trigger:* Group under-supervised due to drift; serious further offence ensues; regulator finds model/threshold gap $\Rightarrow R_t = 1$.
Remedy: Retrain with parity constraints; human-override rule; victim liaison and support ($M_t, J_t^v = 1$).
Confirm: Backtest cohorts; shadow live with subgroup audits; external certification; class closed on pass.

19.10 Dashboards and metrics (ledger view)

- **Harm/recurrence:** assaults; self-harm/suicide; unlawful detention days; recurrence by class; time-to-remedy; time-to-closure T_c^*.
- **Closure:** # classes with $S^* = 1$; median T_c^*; age/backlog of open dockets.
- **Culture (love):** ACCT compliance; safer-cell coverage; programme completion rates; restorative conferences held (ΔL_t proxies).
- **Freedom:** timely releases; successful appeals/complaints; visits/communications fulfilled; recall accuracy (ΔF_t).
- **Rehabilitation outcomes:** post-release education/employment; reoffending for the same index offence (vector-specific).

19.11 Anti-gaming and integrity

- **Suppression control.** Mandatory incident registers; protected disclosures; random wing/team audits; penalties for under-reporting.
- **Scope creep.** Lock class definitions; do not relabel repeats as "new" unless the causal vector differs.
- **Vendor accountability.** For EM and risk tools, require artefacts and SLAs in confirmation packs; no black-box exemptions.

- **Equity checks.** Parity analysis by cohort (age, disability, protected characteristics) in confirmation tests.

19.12 One-page checklist (drop-in for prisons & probation)

Corrections Incident → Confirmation Checklist

Trigger captured? case IDs, unit/wing/team, dates, artefacts
Gate set? caseload reduction; senior sign-off; hard-gates for risky decisions
Docket opened? root cause, fix, tests, owner & deadline, publish plan
Remedy enacted? release/credit; clinical remediation; compensation/support; accountability
Tests passed? case-file replay ✓ stress scenarios ✓ unannounced audits ✓
Scope locked? class definition frozen; distinct vectors filed as c'
Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined
Gates removed? only post-closure; rationale logged

19.13 Limits and open problems

- **Attribution.** Multi-actor causality (governor, healthcare, probation, vendor) complicates C_t; define shared-fault taxonomies.
- **Data quality.** Incomplete logs and conflicting records impede evidence; invest in auditable data capture.
- **Rare but severe events.** Low base rates slow T_c^*; rely on high-fidelity drills and external audits while preserving honesty locks.

19.14 What this chapter contributes

A full translation of the minimal-trigger doctrine to corrections: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and evidence standards. With typal/concave credits and $\kappa > 0$, repeating the same vector adds cost without adding kind-level value; one-off and confirmed closure dominates.

Chapter 20

NHS: Sentinel Events, Dockets, and Structural Closure

Notation

Domain mapping (UK NHS & allied services).
An *event* is a documented clinical/operational occurrence with adjudication potential (e.g., procedure, discharge, transfer, prescribing, triage, data handling).
$R_t \in \{0,1\}$: **rejection** (culpable rights/consent breach or harm $\geq h_{\min}$); $B_t \in \{0,1\}$: breach (e.g., consent failure, safeguarding breach, unlawful data processing); $C_t \in \{0,1\}$: culpability gate (intent/recklessness/gross negligence); $H_t \geq 0$: realized harm (clinical severity scales; "never/sentinel" event).

$M_t, J_t^y \in \{0,1\}$: enacted mercy/justice (disclosure/apology, restitution; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: love/freedom increments (safeguarding done; rights exercised in practice).
Incident classes $c \in C$: causal routes (e.g., wrong-site procedure, LASA medication error, sepsis miss, discharge failure, data leak, safeguarding failure).
Switches: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0,1\}$ per class. Evaluator $\mathcal{J}$ is event-valued (Ch. 9).

20.1 Why an event-valued ledger fits the NHS

Healthcare cultures are rich in *intentions* (policies, trainings) but accountability must ride on *events*: what actually happened to patients, what was repaired, and whether a prior hazardous route is *structurally* closed ($\mathrm{HZ}=0$) and *proven* closed ($S^*=1$). Our ledger avoids (i) over-crediting paper fixes and (ii) tolerating repeated vectors ("we trained again").

20.2 Operational semantics (instantiation)

Rejection R_t (gate).

$R_t = 1 \iff \big(B_t = 1$ (rights/consent/safeguarding/data breach) or $H_t \geq h_{\min}\big)$ an

$C_t = 1$ when an accountable NHS body (trust/ICB/provider) or clinician had adequate knowledge/freedom and crossed a fault threshold

(intent, recklessness, gross negligence). If $C_t = 0$ (accident), record H_t for learning but set $R_t = 0$.

Harm H_t. Use published harm/severity scales and a transparent $h_{\min}$ (e.g., moderate harm, permanent harm, death; never/sentinel events).

Mercy/Justice M_t, J_t^v (event-valued only). $M_t = 1$ when disclosure, apology, restitution, and remedial care are *delivered* to the actually harmed. $J_t^v = 1$ when rectification/accountability is *enacted*: process/tool change, sanctions where appropriate, regulator notifications.

Near-misses and drills. Near-miss ($H_t < h_{\min}$ and no B_t) is *not* $R_t = 1$; convert into an *eventized drill* ($D_t = 1$) that safely reenacts the vector. Drills support confirmation but do not mint M, J^v credits.

20.3 Typical incident classes (c) with scope locks

- **WRONG-SITE** — wrong-site/wrong-patient/wrong-procedure. *Scope:* WHO checklist, marking, time-out, identity checks.
- **LASA** — look-alike/sound-alike medication error. *Scope:* prescribing UI, barcode scanning, pharmacy checks.

- **SEPSIS-MISS** — failure to escalate/treat sepsis. *Scope:* observation thresholds, early warning scores, staffing.
- **DISCH-FAIL** — unsafe discharge/failed handover. *Scope:* criteria-to-reside, meds reconciliation, GP/CMHT handoff.
- **SAFEGUARD** — safeguarding breach (adult/child). *Scope:* referral gates, multi-agency process, flagging.
- **DATA-LEAK** — PHI breach (system/integration/process). *Scope:* DPIA, role access, vendor interfaces.

A *scope lock* freezes inputs, tools, thresholds, and workflow preconditions; new causal vectors open a new class c' rather than relabeling.

20.4 Evaluator and constraints (NHS view; broken to fit)

$$\mathcal{J}_{\text{nhs}} = \sum_{t\geq 0}\Big(\alpha\,\Delta L_t + \gamma\,\Delta F_t + \mu\,M_t + \nu\,J_t^{v}\Big) - \sum_{t\geq 0}\Big(\beta\,H_t + \kappa\,R_t\Big), \tag{20.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau\leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^{*} \in \{0,1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t credits *delivered* quality/safety actions (e.g., staffing fix implemented, checklist compliance verified, drills run with artefacts); ΔF_t credits *exercised* patient rights (consent captured, data access/correction, second opinions).

20.5 Minimal-trigger doctrine (trust policy)

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (human factors/UI/process/staffing), fix, tests, owner+deadline, publication plan (privacy-preserving).

2. **Remedy:** Enact M_t (duty of candour, apology, remedial care, restitution) and J_t^v (rectification/accountability: policy/tool change, training, sanctions if needed).

3. **Confirm:** Reenactment with stronger incentives: retrospective replay $\rightarrow$ sim/manikin $\rightarrow$ live supervised drills or shadow operation. On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats of c add cost (κ and H_t) but no new kind-level credit (typal/concave); dominated by one-off + closure (Ch. 11).

20.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Incident reports; morbidity/mortality review; severity scale assignment; discharge summaries; coroners' findings
B_t (breach)	Consent audit failure; safeguarding adjudication; DPIA/IG breach; standards non-compliance
C_t (culpability)	Safety-case gap vs. SOP; foreseeability; rota/staffing logs; change-control minutes
M_t (mercy)	Duty-of-candour documentation; apology/restitution; remedial care delivered
J_t^v (justice)	Tool/process change shipped; checklist/threshold update; sanctions; regulator notice filed
ΔL_t	Completed drills; staffing fixes delivered; equipment/calibration logs; ward safety boards updated
ΔF_t	Rights exercised: consent recorded; data access/rectification fulfilled; second opinions processed
S_c^*	Passed replay + sim/manikin + supervised live drill; artefacts linked (privacy-preserving)

20.7 Confirmation tests (design pattern for trusts)

Step 1: Retrospective replay. Re-run historical cohorts/cases that match the vector (e.g., same theatre list, ward workflow) and show the failure is removed.
Step 2: Simulation/manikin. Run high-fidelity sim (manikin theatre, dispensing UI sandbox) targeting edge cases; publish pass/fail artefacts.
Step 3: Supervised live drill/shadow. Limited live scope with enhanced supervision; pre-registered pass metrics; promotion to normal operations only after success.
Scope lock. Freeze class scope; if later failures differ causally, open c'.

20.8 Tool-gating and interim safeguards after a trigger

Until $S^*(c) = 1$: narrow indications; add double-checks and barcode gates; require senior sign-off; flag patients at-risk; add conservative protocol defaults (e.g., stop-the-line authority). Remove gates only after closure.

20.9 Micro-vignettes (worked examples)

(A) Wrong-site procedure (WRONG-SITE). *Trigger:* Marking/time-out failed; procedure on wrong side; moderate harm

($H_t \geq h_{\min}$); governance finds SOP gap $\Rightarrow R_t = 1$.
Remedy: Duty of candour, remedial care, compensation ($M_t = 1$); rectification ($J_t^v = 1$): enhanced time-out with hard-stop, barcode+photo check, role assignment.
Confirm: Theatre sim $\rightarrow$ supervised lists with observers; zero recurrences over pre-set cases $\Rightarrow S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) LASA medication error (LASA). *Trigger:* Look-alike drug dispensed; patient harmed; foreseeable UI risk $\Rightarrow C_t = 1,\ R_t = 1$.
Remedy: Disclosure/apology; restitution; UI tall-man lettering; barcode scanning; pharmacist double-check ($M_t, J_t^v = 1$).
Confirm: Sandbox UI tests for confusable pairs; ward drills; shadow live; closure on pass.

(C) Unsafe discharge (DISCH-FAIL). *Trigger:* High-risk patient discharged without meds reconciliation; readmission harm; process gap confirmed $\Rightarrow R_t = 1$.
Remedy: Patient support; pathway fix: discharge checklist, GP handoff by T+24h, pharmacy sign-off ($M_t, J_t^v = 1$).
Confirm: Case replay across wards; stress test at weekend/rota gaps; live audits; class closed on pass.

20.10 Dashboards and metrics (ledger view)

- **Harm/recurrence:** never/sentinel events; severity-adjusted incident rates; recurrence by class; time-to-remedy; time-to-closure T_c^*.

- **Closure:** # classes with $S^* = 1$; median T_c^*; backlog/age of open dockets.

- **Culture (love):** drills completed; staffing fixes delivered; safety huddles/boards; patient-partnering activities (ΔL_t proxies).

- **Freedom:** consent capture; second-opinion rates; data access/rectification turn-around (ΔF_t).

20.11 Anti-gaming and integrity

- **Suppression control.** Mandatory incident registers; protected disclosures; random chart/telemetry audits; penalties for under-reporting.

- **Scope creep.** Lock class definitions; no relabeling of repeats as "new" unless causal vector differs.

- **Privacy.** Publish artefacts with redactions or enable independent audit when public release risks confidentiality.

- **Equity.** Require subgroup checks (age, disability, protected characteristics) in confirmation packs.

20.12 One-page checklist (drop-in for NHS trusts)

NHS Incident → Confirmation Checklist

Trigger captured? timestamp, location, service line, artefacts

Gate set? double-checks; barcode gates; senior sign-off; stop-the-line

Docket opened? root cause, fix, tests, owner & deadline, publish plan

Remedy enacted? duty-of-candour; remedial care; restitution; accountability

Tests passed? replay ✓ sim/manikin ✓ supervised live ✓ (artefacts linked)

Scope locked? class frozen; distinct vectors filed as c'

Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined

Gates removed? only post-closure; rationale logged

20.13 Limits and open problems

- **Attribution.** Multi-actor causality (clinician, ward, IT vendor, trust board) complicates C_t; adopt shared-fault taxonomies.

- **Rare but severe vectors.** Low base rates slow T_c^*; rely on high-fidelity drills while keeping honesty locks (no M, J^v for

drills).

- **Data quality.** Incomplete logs impede evidence; invest in auditable capture and registries.

20.14 What this chapter contributes

A full translation of the minimal-trigger doctrine to NHS practice: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and evidence standards. With typal/concave credits and $\kappa > 0$, repeating the same vector adds cost without adding kind-level value; one-off and confirmed closure dominates.

Chapter 21

Education: Event-Valued Discipline and Restoration

Notation

Domain mapping (schools, colleges, universities).
An *event* is a policy-relevant educational occurrence (classroom, assessment, safeguarding, data handling, exclusion/discipline) with adjudication potential.
$R_t \in \{0,1\}$: **rejection** (culpable rights/consent breach or harm $\geq h_{\min}$); $B_t \in \{0,1\}$: breach of policy/rights (safeguarding, assessment integrity, data privacy, disability accommodations); $C_t \in \{0,1\}$: culpability (intent/recklessness/gross negligence by an accountable staff/body); $H_t \geq 0$: realized harm (safeguarding harm, educational deprivation, reputational/legal harm).
$M_t, J_t^v \in \{0,1\}$: *enacted* mercy/justice (restoration to

pupils/students/staff; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: love/freedom increments (care delivered; rights exercised in practice).
Incident classes $c \in \mathcal{C}$ (illustrative): **SAFEGUARD** safeguarding breach; **EXAM-INTEG** assessment/proctoring failure; **SEND-ADJ** failure to provide reasonable adjustments; **BULLY** unresolved bullying/harassment; **DATA-PRIV** student data breach; **ATTEND** unlawful exclusion/attendance process failure.
Switches: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0, 1\}$ per class. Evaluator $\mathcal{J}$ is event-valued (Ch. 9).

21.1 Why education needs an event-valued ledger

Educational governance often tallies *intentions* (policies, assemblies, anti-bullying posters) and *capacities* (proctoring systems, pastoral frameworks). We insist on *events*: what actually happened to pupils/students, what was repaired, and whether the prior hazardous route is *structurally* closed (HZ = 0) and *proven* closed ($S^* = 1$). This blocks over-crediting paper fixes and tolerating repeated vectors ("we reminded staff again").

21.2 Operational semantics (instantiation)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \,(\text{rights/policy breach}) \,\text{or}\, H_t \geq h_{\min}\big) \,\text{and}\, C_t = 1.$$

$C_t = 1$ when an accountable actor (school/college/university unit) had adequate knowledge & freedom and crossed a fault threshold (intent, recklessness, gross negligence). If $C_t = 0$ (accident), record H_t for learning but set $R_t = 0$.

Harm H_t. Use published scales: safeguarding (tiers), educational deprivation days (missed provision/exclusion), documented harassment harm, verified data/privacy harm. Publish $h_{\min}$.

Mercy/Justice M_t, J_t^v (event-valued only). $M_t = 1$ when restoration is *delivered* to those actually harmed (support, tutoring/catch-up hours, apology/compensation where appropriate). $J_t^v = 1$ when rectification/accountability is *enacted* (policy/tool/process change, sanctions where needed).

Near-misses and drills. A near-miss (e.g., exam breach prevented by invigilator; bullying intercepted early) is *not* $R_t = 1$; log and create an *eventized drill* ($D_t = 1$) that safely reenacts the vector. Drills support S^* but do not mint M, J^v.

21.3 Evaluator and constraints (education view; broken to fit)

$$\mathcal{J}_{\text{edu}} = \sum_{t \geq 0} \Big(\alpha\, \Delta L_t + \gamma\, \Delta F_t + \mu\, M_t + \nu\, J_t^{v} \Big) - \sum_{t \geq 0} \Big(\beta\, H_t + \kappa\, R_t \Big), \tag{21.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau \leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^{*} \in \{0,1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t credits *delivered* support (safeguarding actions completed, tutoring hours delivered, restorative meetings held); ΔF_t credits *exercised* rights (reasonable adjustments, appeals upheld, access restored).

21.4 Minimal-trigger doctrine for education

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class c $\Rightarrow$ open a *remediation docket*: root cause (supervision/curriculum/timetable/process/IT), fix, tests, owner+deadline, publication plan (privacy-preserving).

2. **Remedy:** Enact M_t to those harmed (support, catch-up provision, record correction) and J_t^{v} (rectification/accountability: process/tool changes, staff training sanctions if needed).

3. **Confirm:** Reenactment with stronger incentives: retrospective replay → scenario drills → monitored operation. On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats of c add cost (κ and H_t) but no new kind-level credit (typal/concave); dominated by one-off + closure (Ch. 11).

21.5 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Safeguarding/incident reports; exclusion days; missed provision logs; documented harassment outcomes
B_t (breach)	Findings from safeguarding/academic integrity/data protection processes; equality/adjustment audit fails
C_t (culpability)	Safety-case/process gap vs. policy; foreseeability; governance minutes; staff duty records
M_t (mercy)	Support/counselling delivered; tutoring hours; timetable adjustment; apology/compensation; record corrections
J_t^v (justice)	Policy/tool/process change shipped; proctoring redesign; sanctions; appeals remedies
ΔL_t	Restorative meetings; safeguarding actions completed; staff deployment fixes; parent/student engagement logs
ΔF_t	Adjustments delivered (SEND/ADA); appeal outcomes; re-admission/access restored; assessment resits
S_c^*	Passed replay + drills + monitored operation; artefacts logged (privacy-preserving)

21.6 Incident classes & scope locks (examples)

- **SAFEGUARD** — safeguarding breach. *Scope:* referral gates, supervision ratios, escalation timelines, multi-agency coordination.
- **EXAM-INTEG** — assessment misconduct/proctoring failure. *Scope:* invigilation, room layout, remote proctor configs, item exposure controls.
- **SEND-ADJ** — reasonable adjustments not provided. *Scope:* EHCP/plan mapping, exam adjustments, assistive tech, timetabling.
- **BULLY** — unresolved bullying/harassment. *Scope:* reporting routes, mediation/restorative practice, supervision in hotspots.
- **DATA-PRIV** — unlawful data processing/disclosure. *Scope:* DPIA, access controls, vendor integrations.
- **ATTEND** — unlawful exclusion/attendance handling. *Scope:* thresholds, review panels, alternative provision.

21.7 Confirmation tests (design pattern for schools/colleges)

Step 1: Retrospective replay. Re-run case types that match the vector (e.g., similar cohorts/exams/timetables) and show the failure is removed.

Step 2: Scenario drills. High-fidelity safeguarding drills; exam-room simulations; harassment response walk-throughs; red-team item exposure.
Step 3: Monitored operation. Limited live scope under enhanced supervision (random spot-checks, independent observers) with pre-registered pass metrics.
Scope lock. Freeze class scope; new causal vectors open c' rather than relabeling.

21.8 Tool-gating and interim safeguards after a trigger

Until $S^*(c) = 1$: restrict risky assessments (format/rooming), add supervision, hard-gate adjustments (no assessment without confirmed accommodations), quarantine insecure data flows, implement conservative defaults (e.g., corridor duty rosters). Remove gates only after closure.

21.9 Micro-vignettes (worked examples)

(A) Exam integrity breach (EXAM-INTEG). *Trigger:* Item exposure via device in exam hall; cohort disadvantage; fault confirmed $\Rightarrow R_t = 1$.
Remedy: Resit opportunity or moderated grades; apology/support ($M_t = 1$); proctoring redesign, signal blocking, seating plan, device check-in ($J_t^y = 1$).
Confirm: Replay similar exams; drills with invigilators; monitored

exams with observers; pass $\Rightarrow S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Failure to provide adjustments (SEND-ADJ). *Trigger:* Student lacked mandated extra time/assistive tech; learning loss ($H_t \geq h_{\min}$); process gap $\Rightarrow R_t = 1$.
Remedy: Make-up provision; adjusted grading/assessment window; tech installed; staff training ($M_t, J_t^y = 1$).
Confirm: Case replay across the cohort; timetable audit scripts; monitored assessments confirming delivery; closure on pass.

(C) Safeguarding breach (SAFEGUARD). *Trigger:* Supervision failure in known hotspot; bullying incident causes harm; culpability found $\Rightarrow R_t = 1$.
Remedy: Support to harmed; restorative conference; supervision plan; hotspot redesign; sanctions as needed ($M_t, J_t^y = 1$).
Confirm: Break-time drills; supervised periods with spot-checks; incident rates drop; class closed on pass.

21.10 Dashboards and metrics (ledger view)

- **Harm/recurrence:** safeguarding incidents; exam breaches; missed-adjustment counts; recurrence by class; time-to-remedy; T_c^*.

- **Closure:** # classes with $S^* = 1$; median T_c^*; backlog/age of open dockets.

- **Culture (love):** support sessions delivered; restorative meetings; supervision coverage; parent/student engagement (ΔL_t proxies).
- **Freedom:** adjustments delivered; appeals upheld; access restored; fair assessment opportunities (ΔF_t).

21.11 Anti-gaming and integrity

- **Suppression control.** Mandatory incident registers; protected disclosures; random audits of exam rooms/timetables/logs.
- **Scope creep.** Lock class definitions; do not relabel repeats as "new" unless causal vector differs.
- **Equity checks.** Subgroup analysis (SEND, age, protected characteristics) in confirmation packs.
- **Transparency.** Publish artefacts with privacy safeguards or allow independent audit.

21.12 One-page checklist (drop-in for schools/colleges)

Education Incident → Confirmation Checklist

Trigger captured? time, cohort, room/timetable, artefacts
Gate set? risky assessments constrained; supervision added; adjustments hard-gated; data flows quarantined

Docket opened? root cause, fix, tests, owner & deadline, publish plan
Remedy enacted? support/catch-up; apology/compensation where apt; record correction; accountability
Tests passed? replay ✓ scenario drills ✓ monitored operation ✓
Scope locked? class frozen; distinct vectors filed as c'
Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined
Gates removed? only post-closure; rationale logged

21.13 Limits and open problems

- **Attribution.** Multi-actor causality (teacher, exams office, IT, duty staff) complicates C_t; define shared-fault taxonomies.
- **Measurement.** Educational deprivation is partly latent; pair quantitative logs with qualitative panels.
- **Rare but impactful events.** Some vectors are infrequent; rely on high-fidelity drills while keeping honesty locks.

21.14 What this chapter contributes

A full translation of the minimal-trigger doctrine to education: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and evidence standards. With typal/concave credits and $\kappa > 0$,

repeating the same vector adds cost without adding kind-level value; one-off and confirmed closure dominates.

Chapter 22

Adult Social Care: Safeguarding, Dignity, and Confirmation

Notation

Domain mapping (adult social care: domiciliary, residential, supported living).

An *event* is a care interaction or provider action/inaction with adjudication potential.

$R_t \in \{0,1\}$: **rejection** (culpable rights/consent breach or harm $\geq h_{\min}$); $B_t \in \{0,1\}$: breach (safeguarding, consent/capacity, unlawful restriction, financial abuse, data/privacy); $C_t \in \{0,1\}$: culpability gate (intent/reckless/gross negligence by provider/commissioner); $H_t \geq 0$: realized harm (injury, neglect,

nutritional/hydration harm, pressure injury, dignity breach, financial loss).
$M_t, J_t^y \in \{0,1\}$: enacted mercy/justice (restoration to the actually harmed; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: love/freedom increments (safeguarding actions delivered; rights exercised in practice).
Incident classes $c \in C$ (illustrative): **MISS-VISIT** missed/shortened domiciliary visits; **MED-ADMIN** medication administration failure; **PRESSURE** preventable pressure injury; **HYDRATION** dehydration/malnutrition; **FIN-ABUSE** financial exploitation; **CAPACITY** consent/capacity breach; **UNLAWF-RESTR** unlawful restraint/deprivation of liberty; **DATA-PRIV** data/privacy breach.
Switches: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0,1\}$ per class. Evaluator $\mathcal{J}$ is event-valued (Ch. 9).

22.1 Why adult social care fits an event-valued ledger

Care quality is not a policy binder; it is meals delivered, medication administered, pressure areas protected, and dignities respected *in fact*. We therefore score *events*: what actually happened to the person, whether remedy was enacted, and whether the prior causal route is *structurally* and *provably* closed ($\mathrm{HZ} = 0$, $S^* = 1$). This prevents over-crediting paperwork and tolerating repeated vectors ("we re-trained" without change).

22.2 Operational semantics (instantiation)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \,(\text{rights/safeguarding breach}) \text{ or } H_t \geq h_{\min}\big) \text{ and } C_t = 1.$$

$C_t = 1$ when an accountable provider/commissioner had adequate knowledge & freedom and crossed a fault threshold (intent, recklessness, gross negligence) relative to a published care plan/safety case. If $C_t = 0$ (accident), record H_t for learning but do not set $R_t = 1$.

Harm H_t. Use published scales: clinical harm tiers; grade of pressure injury; hydration/nutrition risk; verified financial loss; dignity/privacy harms. Publish $h_{\min}$.

Mercy/Justice M_t, J_t^v. $M_t = 1$ when restoration is *delivered* to those harmed (remedial care, restitution/compensation, apology, replacement services). $J_t^v = 1$ when rectification/accountability is *enacted* (staffing/rota fix, supervision/gates, sanctions, public notice/regulator filing).

Near-misses and drills. Near-miss (e.g., late visit caught by digital check-in; medication caught before administration) is not $R_t = 1$; convert into an *eventized drill* ($D_t = 1$) for reenactment. Drills support S^* but do not mint M, J^v.

22.3 Evaluator and constraints (care view; broken to fit)

$$\mathcal{J}_{\text{asc}} = \sum_{t\geq 0}\Big(\alpha\,\Delta L_t + \gamma\,\Delta F_t + \mu\,M_t + \nu\,J_t^{v}\Big) - \sum_{t\geq 0}\Big(\beta\,H_t + \kappa\,R_t\Big), \tag{22.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau\leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^{*} \in \{0,1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t credits *delivered* safeguards (pressure-relief schedules met, hydration plans delivered, safeguarding visits completed); ΔF_t credits *exercised* freedoms/rights (choices respected, access to visitors/advocates, lawful restrictions with authorization and review).

22.4 Minimal-trigger doctrine (commissioner/provider policy)

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class c $\Rightarrow$ open a *remediation docket*: root cause (rostering/IT/transport/staffing/capacity/oversight), fix, tests, owner+deadline, publication plan (privacy-preserving).
2. **Remedy:** Enact M_t (make-up care, restitution/compensation, apology, clinical remediation) and

J_t^v (rectification/accountability: rota redesign, staffing floor, escalation gates, sanctions if needed).

3. **Confirm:** Reenactment with stronger incentives: replay → stress drills (peak rosters, weather/transport shocks) → monitored live. On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats of c add cost (κ and H_t) but no new kind-level credit (typal/concave); dominated by one-off + closure (Ch. 11).

22.5 Incident classes & scope locks (examples)

- **MISS-VISIT** — missed/shortened domiciliary visits. *Scope:* e-rostering, travel buffers, check-in/checkout, lone-worker policy, escalation when late.

- **MED-ADMIN** — medication administration failure. *Scope:* MAR charts, double-checks, pharmacy liaison, training/competency.

- **PRESSURE** — preventable pressure injury. *Scope:* risk assessment, turning schedules, equipment availability, audit.

- **HYDRATION** — dehydration/malnutrition. *Scope:* fluid/meal plans, recording, mealtime support, shopping/meal delivery logistics.

- **FIN-ABUSE** — financial abuse/theft. *Scope:* cash handling, shopping protocols, CCTV/logs, two-person rule.

- **CAPACITY** — consent/capacity breach. *Scope:* assessments, best-interest decisions, advocacy/IMCA, review cadence.

- **UNLAWF-RESTR** — unlawful restraint/deprivation of liberty. *Scope:* legal authorization pathways, observation logs, proportionality tests, review.

- **DATA-PRIV** — data/privacy breach. *Scope:* DPIA, role-based access, device policy, vendor links.

Each class must have a written *scope lock* to prevent relabeling.

22.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Clinical incident reports; hydration/weight loss logs; pressure injury grades; verified financial loss; dignity/privacy breach records
B_t (breach)	Safeguarding findings; consent/capacity assessment failures; unlawful restriction determinations; DPIA/IG breaches
C_t (culpability)	Safety-case/process gap vs. plan; foreseeability; rota/staffing logs; training/competency evidence
M_t (mercy)	Remedial care delivered; make-up hours; restitution/compensation; apology; advocacy access records
J_t^v (justice)	Rota/tool/process change shipped; staffing floors met; sanctions/contract actions; regulator/commissioner notice
ΔL_t	Completed safeguarding actions; pressure-relief compliance; hydration program adherence; family/advocate engagement
ΔF_t	Choices exercised (meals, routines); visitors/advocates access; lawful restriction reviews; appeals upheld
S_c^*	Passed replay + stress drills + monitored live; artefacts recorded (privacy-preserving)

22.7 Confirmation tests (design pattern for care)

Step 1: Replay. Re-run comparable caseload weeks/routes; show removal of prior vector (e.g., no missed visits under similar travel loads).
Step 2: Stress drills. Simulate peak sickness/rota gaps, weather disruption, pharmacy delays; confirm gates hold.
Step 3: Monitored live. Limited live window with independent spot-checks (e-check-in data, family calls, random home visits); pre-registered pass metrics.
Scope lock. Freeze class scope; new causal vectors open c'.

22.8 Tool-gating and interim safeguards after a trigger

Until $S^*(c) = 1$: add call-ahead buffers; require supervisor sign-off for shortened visits; enable visit-verification tech; prioritize high-risk clients; temporary two-person visits for HIGH risk; pause non-essential tasks. Remove gates only after closure.

22.9 Micro-vignettes (worked examples)

(A) Missed domiciliary visits (MISS-VISIT). *Trigger:* Multiple missed evening calls; dehydration/skin breakdown ensues ($H_t \geq h_{\min}$); rostering gap confirmed $\Rightarrow R_t = 1$.
Remedy: Make-up care; restitution; hydration/pressure plans; rota

redesign with travel buffers; escalation script ($M_t, J_t^y = 1$).
Confirm: Replay prior routes; stress drill at peak sickness; monitored live with e-check-ins $\Rightarrow S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Consent/capacity breach (CAPACITY). *Trigger:* Restrictive practice used without lawful authorization/review; dignity harm; governance finds process gap $\Rightarrow R_t = 1$.
Remedy: Apology; advocacy; immediate review and proper authorization pathway; staff retraining; audit trail ($M_t, J_t^y = 1$).
Confirm: Case-file replay; stress tests on night/weekend coverage; random audits; closure on pass.

(C) Financial abuse (FIN-ABUSE). *Trigger:* Carer misuses client funds; verified loss ($H_t \geq h_{\min}$) and breach ($B_t = 1$).
Remedy: Full restitution; police/referral as appropriate; two-person cash rule; receipts/audit upgrades ($M_t, J_t^y = 1$).
Confirm: Shadow audits with test purchases; spot-checks; no recurrences $\Rightarrow$ class closed.

22.10 Dashboards and metrics (ledger view)

- **Harm/recurrence:** missed/shortened visit rates; pressure injury incidence; hydration/weight alerts; recurrence by class; time-to-remedy; time-to-closure T_c^*.

- **Closure:** # classes with $S^* = 1$; median T_c^*; age/backlog of open dockets.

- **Culture (love):** safeguarding actions completed; family/advocate engagement; staff supervision rates (ΔL_t proxies).

- **Freedom:** rights exercised (choices honored, advocate access); lawful restriction reviews passed; upheld appeals (ΔF_t).

22.11 Anti-gaming and integrity

- **Suppression control.** Mandatory incident registers; protected disclosures; random spot-calls with families; penalties for under-reporting.

- **Scope creep.** Lock class definitions; do not relabel repeats as "new" unless the causal vector differs.

- **Vendor accountability.** For e-rostering/verification tools, require artefacts in confirmation packs; no black-box exemptions.

- **Equity checks.** Cohort analysis (age/disability/protected characteristics) in confirmation tests.

22.12 One-page checklist (drop-in for providers/commissioners)

Adult Social Care Incident → Confirmation Checklist

Trigger captured? client IDs, dates/times, rota/IT logs, artefacts
Gate set? buffers; supervisor sign-off; visit-verification tech; HIGH-risk two-person rule
Docket opened? root cause, fix, tests, owner & deadline, publish plan
Remedy enacted? make-up care; restitution/compensation; clinical remediation; accountability
Tests passed? replay ✓ stress drills ✓ monitored live ✓
Scope locked? class frozen; distinct vectors filed as c'
Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined
Gates removed? only post-closure; rationale logged

22.13 Limits and open problems

- **Attribution.** Multi-actor causality (provider, commissioner, GP/pharmacy, transport/vendor) complicates C_t; define shared-fault taxonomies.

- **Data quality.** Incomplete logs and paper records impede evidence; invest in auditable capture.

- **Rare but severe events.** Low base rates slow T_c^*; rely on drills and external audits while preserving honesty locks.

22.14 What this chapter contributes

A full translation of the minimal-trigger doctrine to adult social care: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and evidence standards. With typal/concave credits and $\kappa > 0$, repeating the same vector adds cost without adding kind-level value; one-off and confirmed closure dominates.

Chapter 23

Mental Health Services: Crisis, Restraint, and Confirmation

Notation

Domain mapping (mental health: crisis, inpatient, community, CAMHS).
An *event* is a clinically/policy-relevant occurrence with adjudication potential (crisis call, assessment, admission/discharge, restraint/seclusion, medication, safeguarding, data handling). $R_t \in \{0,1\}$: **rejection** (culpable rights/consent breach or harm $\geq h_{\min}$); $B_t \in \{0,1\}$: breach (consent/capacity, unlawful restriction, safeguarding, privacy/data); $C_t \in \{0,1\}$: culpability gate (intent/recklessness/gross negligence by an accountable ser-

vice/clinician); $H_t \geq 0$: realized harm (injury, self-harm/suicide, unlawful detention days, dignity harm).
$M_t, J_t^{\nu} \in \{0, 1\}$: *enacted* mercy/justice (restoration to the actually harmed; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: increments to care/safety (love) and exercised rights/freedoms (freedom).
Incident classes $c \in C$ (illustrative): **CRISIS-DELAY** delayed/failed crisis response; **RESTR-USE** restraint/seclusion outside policy; **LIG-RISK** ligature/observation failure; **DISCH-GAP** unsafe discharge/continuity gap; **MED-PSY** psychotropic medication error; **CAPACITY** consent/capacity breach; **DATA-PRIV** privacy breach.
Switches: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0, 1\}$ per class. Evaluator $\mathcal{J}$ is event-valued (Ch. 9).

23.1 Why an event-valued ledger fits mental health

Mental health outcomes are about *what actually happened*: whether a crisis call was answered in time, whether restrictive practices were lawful and proportionate, whether discharge led to safe continuity of care, and whether repeat routes are *structurally* closed (HZ = 0) and *proven* closed ($S^* = 1$). Counting training hours or policy binders over-credits intentions; the ledger scores events on the realized path.

23.2 Operational semantics (instantiation)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \,(\text{rights/safeguarding breach}) \,\text{or}\, H_t \geq h_{\min}\big) \,\text{and}\, C_t = 1.$$

$C_t = 1$ when an accountable actor (provider/ward/team) had adequate knowledge/freedom and crossed a fault threshold (intent, recklessness, gross negligence). If $C_t = 0$ (accident), record H_t for learning but do not set $R_t = 1$.

Harm H_t. Use published scales: self-harm/suicide, injury tiers, unlawful detention days, restraint/seclusion duration beyond policy, dignity/privacy harm. Publish $h_{\min}$ per service.

Mercy/Justice M_t, J_t^v. $M_t = 1$ when restoration is *delivered* (apology, remedial care, restitution, record correction). $J_t^v = 1$ when rectification/accountability is *enacted* (policy/process/tool change, staffing/supervision fix, sanctions, regulator notice).

Near-misses and drills. Near-miss (e.g., observation lapse caught before harm) is not $R_t = 1$; convert to *eventized drills* ($D_t = 1$). Drills support S^* but do not mint M, J^v.

23.3 Evaluator and constraints (mental health view; broken to fit)

$$\mathcal{J}_{\text{mh}} = \sum_{t\geq 0}\Big(\alpha\,\Delta L_t + \gamma\,\Delta F_t + \mu\,M_t + \nu\,J_t^{v}\Big) - \sum_{t\geq 0}\Big(\beta\,H_t + \kappa\,R_t\Big), \tag{23.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau\leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^{*} \in \{0,1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t credits *delivered* care/safety actions (observation levels met, crisis lines staffed, de-escalation drills done); ΔF_t credits *exercised* rights (lawful consent, advocacy access, appeals/tribunals).

23.4 Minimal-trigger doctrine (service policy)

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (staffing/rota, observation, care plan, environment, IT), fix, tests, owner+deadline, publication plan (privacy-preserving).

2. **Remedy:** Enact M_t to those harmed (apology, remedial care, restitution, record corrections) and J_t^{v} (rectification/accountability: policy change, staffing floor, environment fix, sanctions if needed).

3. **Confirm:** Reenactment with stronger incentives: ret-

rospective replay → high-fidelity sims → monitored live. On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats of c add cost (κ and H_t) but no new kind-level credit (typal/concave); dominated by one-off + closure (Ch. 11).

23.5 Incident classes & scope locks (examples)

- **CRISIS-DELAY** — crisis line/response delay leading to harm. *Scope:* staffing, overflow routing, escalation, inter-agency handoffs.

- **RESTR-USE** — restraint/seclusion outside policy (thresholds, duration, documentation). *Scope:* de-escalation, environment, observation, authorisation.

- **LIG-RISK** — ligature/observation failure (environmental risks; observation levels not met). *Scope:* environment checks, rota, monitoring tech, handoffs.

- **DISCH-GAP** — unsafe discharge/continuity-of-care gap. *Scope:* meds reconciliation, follow-up within target window, GP/CMHT handoff, social support linkage.

- **MED-PSY** — psychotropic medication error (dose/drug/interactions). *Scope:* prescribing/administration, double-checks, pharmacy liaison.

- **CAPACITY** — consent/capacity breach or unlawful restriction. *Scope:* assessments, advocacy, review cadence, documentation.

- **DATA-PRIV** — privacy/data breach (notes, portals, devices, third-party apps). *Scope:* DPIA, role access, vendor links.

Each class must include a *scope lock* freezing inputs, thresholds, tools, and preconditions; new causal vectors open a new class c'.

23.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Incident reports; self-harm/suicide logs; injury reports; unlawful detention days; dignity/privacy harm records
B_t (breach)	Safeguarding findings; consent/capacity adjudication; unlawful restriction determinations; privacy/IG breaches
C_t (culpability)	Safety-case/process gap vs. policy; foreseeability; rota/staffing logs; handoff/observation records
M_t (mercy)	Remedial care delivered; apology; restitution/compensation; record corrections; advocacy access
J_t^v (justice)	Policy/process/environment change; staffing floors; sanctions; regulator/ombudsman notice
ΔL_t	Observation compliance; crisis-line SLA delivery; de-escalation drills; environment risk fixes completed
ΔF_t	Advocacy/tribunal access; consent properly recorded; appeals upheld; lawful restriction reviews
S_c^*	Passed replay + sim + monitored live; artefacts linked (privacy-preserving)

23.7 Confirmation tests (design pattern for MH services)

Step 1: Retrospective replay. Re-run matched cohorts: crisis-call volumes, ward observation intervals, discharge pathways; show removal of the prior vector.
Step 2: Simulation. High-fidelity ward sims (de-escalation, observation load), crisis-call load tests, environment checks (ligature sweeps); publish pass/fail artefacts.
Step 3: Monitored live. Limited live window with enhanced supervision (random spot checks, independent observers) and pre-registered pass metrics.
Scope lock. Freeze class scope; new causal vectors require a new class c'.

23.8 Tool-gating and interim safeguards after a trigger

Until $S^*(c) = 1$: increase observation levels; install interim environment fixes (remove ligature points, safe rooms); add staffing; require senior approval for restrictive practices; fast-track advocacy access; conservative defaults on discharge (follow-up within 24–72h). Remove gates only after closure.

23.9 Micro-vignettes (worked examples)

(A) Restraint outside policy (RESTR-USE). *Trigger:* Prolonged seclusion without review; injury occurs; documentation shows threshold not met $\Rightarrow B_t = 1, C_t = 1$, so $R_t = 1$.
Remedy: Apology, remedial care, record correction ($M_t = 1$); rectification ($J_t^v = 1$): staffing floor, de-escalation training, environment change, authorisation gates.
Confirm: Ward sims with independent observers; monitored live period; no policy breaches under stress; $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Ligature risk/observation failure (LIG-RISK). *Trigger:* Observation level not maintained; attempt occurs; coroner/ombudsman finds gross failure $\Rightarrow R_t = 1$.
Remedy: Family support/restitution; environment remediation; rota redesign; monitoring tech; accountability ($M_t, J_t^v = 1$).
Confirm: Environment sweeps; load tests of observation rotas; surprise audits; class closed on pass.

(C) Unsafe discharge (DISCH-GAP). *Trigger:* Discharge without meds reconciliation or follow-up; relapse and harm; process gap confirmed $\Rightarrow R_t = 1$.
Remedy: Outreach and remedial care; discharge checklist; follow-up window; GP/CMHT handoff; social support linkage ($M_t, J_t^v = 1$).
Confirm: Case replay on similar discharges; stress test at weekends/rota gaps; monitored live; closure on pass.

23.10 Dashboards and metrics (ledger view)

- **Harm/recurrence:** self-harm/suicide rates; restraint/seclusion outside policy; recurrence by class; time-to-remedy; time-to-closure T_c^*.
- **Closure:** # classes with $S^* = 1$; median T_c^*; backlog/age of open dockets.
- **Culture (love):** observation compliance; de-escalation drills; environment fixes; family/advocacy engagement (ΔL_t proxies).
- **Freedom:** advocacy access; tribunal/appeal timeliness; lawful restriction reviews passed; consent capture (ΔF_t).

23.11 Anti-gaming and integrity

- **Suppression control.** Mandatory incident registers; protected disclosures; random ward/audit checks; penalties for under-reporting.
- **Scope creep.** Lock class definitions; no relabeling of repeats as "new" unless causal vector differs.
- **Privacy.** Publish artefacts with redactions or enable independent audit where public release risks confidentiality.
- **Equity.** Cohort analysis (age, sex, disability, protected characteristics) in confirmation packs.

23.12 One-page checklist (drop-in for MH providers)

Mental Health Incident → Confirmation Checklist

Trigger captured? timestamps, ward/team, artefacts
Gate set? higher observation; environment fixes; senior approval for restrictive practices
Docket opened? root cause, fix, tests, owner & deadline, publish plan
Remedy enacted? apology/remedial care; restitution; advocacy access; accountability
Tests passed? replay ✓ simulation ✓ monitored live ✓
Scope locked? class frozen; distinct vectors as c'
Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined
Gates removed? only post-closure; rationale logged

23.13 Limits and open problems

- **Attribution.** Multi-actor causality (ward, crisis team, community services, social care, vendors) complicates C_t; adopt shared-fault taxonomies.

- **Data quality.** Incomplete observation/handoff logs impede evidence; invest in auditable capture and environment telemetry.

- **Rare but severe events.** Low base rates slow T_c^*; rely on high-fidelity drills and external audits while preserving honesty locks.

23.14 What this chapter contributes

A full translation of the minimal-trigger doctrine to mental health services: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and evidence standards. With typal/concave credits and $\kappa > 0$, repeating the same vector adds cost without adding kind-level value; one-off and confirmed closure dominates.

Chapter 24

Housing & Homelessness: Safety, Repairs, and Confirmation

Notation

Domain mapping (local authorities, housing associations, private landlords, TA providers, homelessness services). An *event* is a housing/homelessness outcome with adjudication potential (hazard, inspection, repair, placement, eviction, complaint, data handling).

$R_t \in \{0,1\}$: **rejection** (culpable rights/consent breach or harm $\geq h_{\min}$); $B_t \in \{0,1\}$: breach (e.g., unlawful eviction, HHSRS Cat. 1 hazard unremedied, fire/gas/electrical noncompliance, gatekeeping); $C_t \in \{0,1\}$: culpability gate (in-

tent/reckless/gross negligence by landlord/authority/provider); $H_t \geq 0$: realized harm (injury/illness, unsafe exposure, loss of home/rights, verified financial loss).
$M_t, J_t^{\nu} \in \{0, 1\}$: *enacted* mercy/justice (restoration to the actually harmed; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: increments to care/safety (love) and exercised rights/freedoms (freedom).
Incident classes $c \in C$ (illustrative): **HHSRS-HAZ** Category 1 hazard; **DAMP-MOULD**; **FIRE-SAFE**; **GAS-ELEC**; **REPAIR-FAIL**; **TEMP-UNSAFE** (temporary accommodation); **EVICT-DUEPROC** unlawful/retaliatory eviction; **HOMELESS-GATE** gatekeeping/decision error; **ASB-RESP** failure to act on anti-social behaviour; **DATA-PRIV** tenant data breach.
Switches: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0, 1\}$ per class. Evaluator $\mathcal{J}$ is event-valued (Ch. 9).

24.1 Why an event-valued ledger fits housing

Housing outcomes are concrete: hazards are removed (or not), repairs are completed (or not), families are safely accommodated (or not). Counting policies or response *intentions* over-credits paperwork. We therefore score *events on the realized path* and require proof that prior hazardous routes are *structurally* and *provably* closed (HZ = 0, $S^* = 1$).

24.2 Operational semantics (instantiation)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \,(\text{rights/safety breach}) \,\text{or}\, H_t \geq h_{\min}\big) \,\text{and}\, C_t = 1.$$

$C_t = 1$ when an accountable actor (landlord/ALMO/housing association/local authority/provider) had adequate knowledge & freedom and crossed a fault threshold (intent, recklessness, gross negligence) relative to a published safety case (HHSRS/Fire/Gas/Electrical/Allocations). If $C_t = 0$ (accident), record H_t but do not set $R_t = 1$.

Harm H_t. Illness/injury from damp/mould/cold; fire/gas/electrical incidents; homelessness nights; unlawful eviction/lockout; financial loss due to maladministration; dignity/privacy harm. Publish $h_{\min}$.

Mercy/Justice M_t, J_t^v. $M_t = 1$ when remediation is *delivered* to those actually harmed (decant/temporary accommodation, repairs completed, restitution/compensation, apology). $J_t^v = 1$ when rectification/accountability is *enacted* (policy/process/tool change; enforcement; sanctions; regulator/ombudsman engagement).

Near-misses and drills. A near-miss (e.g., alarm caught a fire-risk before harm; unlawful eviction halted) is not $R_t = 1$; log and convert to *eventized drills* ($D_t = 1$). Drills support S^* but do not mint M, J^v.

24.3 Evaluator and constraints (housing view; broken to fit)

$$\mathcal{J}_{\text{hh}} = \sum_{t\geq 0}\Big(\alpha\,\Delta L_t + \gamma\,\Delta F_t + \mu\,M_t + \nu\,J_t^{v}\Big) - \sum_{t\geq 0}\Big(\beta\,H_t + \kappa\,R_t\Big), \tag{24.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau\leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^{*} \in \{0,1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t credits *delivered* safety/care (repairs closed, decants, risk mitigations, inspections completed); ΔF_t credits *exercised* rights (appeals upheld, unlawful eviction reversed, access to waiting-list/statutory duties).

24.4 Minimal-trigger doctrine (landlord/authority policy)

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (fabric, ventilation, heating, access, contractor, decision-making), fix, tests, owner+deadline, publication plan (privacy-preserving).

2. **Remedy:** Enact M_t (decant or make-safe; complete repairs; restitution/compensation; apology) and J_t^{v} (rectification/accountability: policy/process/spec change;

contractor sanctions; regulator notice).

3. **Confirm:** Reenactment under stronger incentives: cohort replay → stress season (cold/wet) → independent inspection/tenant verification. On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats of c add cost (κ and H_t) but no new kind-level credit (typal/concave); dominated by one-off + closure (Ch. 11).

24.5 Incident classes & scope locks (examples)

- **HHSRS-HAZ** — Category 1 hazard not made safe in time. *Scope:* HHSRS assessment, works orders, decant triggers, monitoring.
- **DAMP-MOULD** — damp/mould causing ill-health. *Scope:* building fabric, ventilation, heating, response SLAs, contractor quality.
- **FIRE-SAFE** — fire safety failures (alarms, compartmentation, cladding, egress). *Scope:* FRA actions, waking watch, drills.
- **GAS-ELEC** — expired/failed gas/electrical safety (CP12/EICR), CO incidents. *Scope:* certification cadence, access, shut-off rules.
- **REPAIR-FAIL** — statutory repair failures causing harm. *Scope:* prioritisation, parts, access, contractor management.

- **TEMP-UNSAFE** — unsafe/unsuitable temporary accommodation/overcrowding. *Scope:* placement standards, inspections, length-of-stay caps.

- **EVICT-DUEPROC** — unlawful/retaliatory eviction/lockout. *Scope:* notice validity, court order, protection of goods, homelessness duty.

- **HOMELESS-GATE** — gatekeeping/decision error under homelessness law. *Scope:* eligibility, priority need, interim duty, reviews.

- **ASB-RESP** — failure to act on anti-social behaviour leading to harm. *Scope:* thresholds, joint working, injunctions, support pathways.

- **DATA-PRIV** — tenant data/privacy breach. *Scope:* DPIA, role-based access, vendor links.

Each class must include a written *scope lock* (inputs, standards, SLAs, tools, preconditions); new vectors open c' rather than relabeling.

24.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Medical reports/GP letters; injury logs; homelessness nights; verified financial loss; environmental readings (humidity/CO/temperature)
B_t (breach)	HHSRS/FRA findings; gas/electrical certificates; unlawful eviction findings/court orders; ombudsman decisions
C_t (culpability)	Safety-case/process gap vs. standard; foreseeability; contractor oversight records; governance minutes
M_t (mercy)	Decant/make-safe delivered; repairs completed; alternative accommodation; restitution/compensation; apology
J_t^v (justice)	Policy/spec/process change shipped; enforcement/sanctions; regulator notification; contractor remedies
ΔL_t	Completed inspections; repairs closed within SLA; protective kit (dehumidifiers/heaters) provided; safety comms to tenants
ΔF_t	Appeals/reviews upheld; unlawful evictions reversed; access to duty lists/services; privacy rights fulfilled
S_c^*	Passed cohort replay + seasonal stress + independent inspection/tenant verification (artefacts linked)

24.7 Confirmation tests (design pattern for housing)

Step 1: Cohort replay. Re-run comparable cases (same archetype/blocks/stock type) and show removal of prior vector (e.g., moisture/temperature profiles, repair closure).
Step 2: Seasonal stress. Test during cold/wet months or simulate with forced ventilation/heating patterns; include access/contractor load shocks.
Step 3: Independent inspection & tenant verification. Environmental health/fire service/third-party survey; random tenant callbacks/photos; pre-registered pass metrics.
Scope lock. Freeze class scope; later failures with different causality open c'.

24.8 Tool-gating and interim safeguards after a trigger

Until $S^*(c) = 1$: decant high-risk households; provide heaters/dehumidifiers; disable unsafe gas/electrical appliances; enforce waking watch/alarms; restrict high-risk areas; priority routing of contractors; hold evictions where process is in question. Remove gates only after closure.

24.9 Micro-vignettes (worked examples)

(A) Damp and mould (DAMP-MOULD). *Trigger:* Child respiratory illness linked to mould; repeated reports; HHSRS Cat. 1 confirmed $\Rightarrow B_t = 1, C_t = 1$, so $R_t = 1$.
Remedy: Decant/make-safe; repairs to fabric/ventilation/heating; restitution; apology ($M_t = 1$); policy/spec change, contractor sanctions, proactive inspections ($J_t^y = 1$).
Confirm: Cohort replay across similar stock; seasonal stress; independent inspection/tenant verification $\Rightarrow S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Unlawful eviction (EVICT-DUEPROC). *Trigger:* Tenant locked out without court order; belongings withheld; rights breach $\Rightarrow R_t = 1$.
Remedy: Reinstatement or alternative accommodation; restitution; record correction; apology ($M_t = 1$); rectification ($J_t^y = 1$): staff training, process gates, sanctions.
Confirm: Case replay (recent evictions), random audits of notice/process; independent legal review; closure on pass.

(C) Fire safety failure (FIRE-SAFE). *Trigger:* Non-functioning alarms/blocked egress; FRA actions overdue; incident occurs $(H_t \geq h_{\min}) \Rightarrow R_t = 1$.
Remedy: Make-safe (alarms, compartmentation fixes, waking watch); restitution; accountability ($M_t, J_t^y = 1$).
Confirm: Alarm/egress tests under load; fire-service inspection; tenant drills; class closed on pass.

24.10 Dashboards and metrics (ledger view)

- **Harm/recurrence:** Category 1 hazards; illness/injury incidents; homelessness nights; recurrence by class; time-to-remedy; time-to-closure T_c^*.
- **Closure:** # classes with $S^* = 1$; median T_c^*; backlog/age of open dockets.
- **Culture (love):** proactive inspections; repairs closed to SLA; tenant engagement; safety comms (ΔL_t proxies).
- **Freedom:** upheld reviews/appeals; unlawful evictions reversed; access to services/waiting lists; privacy rights fulfilled (ΔF_t).

24.11 Anti-gaming and integrity

- **Suppression control.** Mandatory incident registers; tenant whistleblower routes; random property spot-checks; penalties for under-reporting.
- **Scope creep.** Lock class definitions; do not relabel repeats as "new" unless causal vector differs.
- **Contractor accountability.** Require artefacts in confirmation packs; performance bonds/withholds linked to T_c^*; no black-box exemptions.
- **Transparency.** Publish privacy-preserving artefacts or permit independent audit (ombudsman/regulator/tenant panel).

24.12 One-page checklist (drop-in for landlords/authorities)

Housing Incident → Confirmation Checklist

Trigger captured? property IDs/UPRNs, dates, inspection/repair logs, photos/telemetry
Gate set? decant/make-safe; disable unsafe systems; priority contractor routing; hold risky evictions
Docket opened? root cause, fix, tests, owner & deadline, publish plan
Remedy enacted? repairs complete; alternative accommodation; restitution/compensation; accountability
Tests passed? cohort replay ✓ seasonal stress ✓ independent inspection/tenant verification ✓
Scope locked? class frozen; distinct vectors filed as c'
Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined
Gates removed? only post-closure; rationale logged

24.13 Limits and open problems

- **Attribution.** Multi-actor causality (landlord, contractor, managing agent, authority, fire service) complicates C_t; adopt shared-fault taxonomies.
- **Resource constraints.** Budgets/stock condition limit speed;

use T_c^* targets, triage, and transparency to prioritise high-harm vectors.

- **Data quality.** Incomplete repair/inspection logs impede evidence; invest in auditable capture and tenant reporting channels.

24.14 What this chapter contributes

A full translation of the minimal-trigger doctrine to housing & homelessness: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and evidence standards. With typal/concave credits and $\kappa > 0$, repeating the same vector adds cost without adding kind-level value; one-off and confirmed closure dominates.

Chapter 25

Energy & Utilities: Outages, Safety, and Confirmation

Notation

Domain mapping (electricity, gas, water, wastewater; optionally district heat).
An *event* is an operational/customer-facing occurrence with adjudication potential (outage, pressure/flow breach, quality exceedance, safety incident, cyber event, billing/priority failure). $R_t \in \{0,1\}$: **rejection** (culpable rights/safety breach or harm $\geq h_{\min}$); $B_t \in \{0,1\}$: breach (licence/standard breach, unlawful disconnection, quality limit exceedance, black-start noncompliance); $C_t \in \{0,1\}$: culpability gate (intent/reckless/gross

negligence by operator/contractor/vendor); $H_t \geq 0$: realized harm (injury, health/quality harm, vulnerable-customer detriment, economic loss beyond threshold).
$M_t, J_t^y \in \{0, 1\}$: *enacted* mercy/justice (restoration to the actually harmed; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: increments to care/safety (love) and exercised rights/freedoms (freedom).
Incident classes $c \in C$ (illustrative): **GRID-OUT** unplanned outage/blackout; **LOADSHED** mis-prioritised load-shed hitting protected sites; **GAS-LEAK** leak/explosion hazard; **WATER-QUAL** drinking-water exceedance/boil notice; **SEWER-OF** sewage overflow; **CYBER-ICS** ICS/SCADA breach causing unsafe state; **WORK-SAFE** worker/public safety incident; **BILL-ERR** billing error/unlawful disconnection of protected customer.
Switches: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0, 1\}$ per class. Evaluator $\mathcal{J}$ is event-valued (Ch. 9).

25.1 Why an event-valued ledger fits utilities

Reliability and safety are judged by *what happened*: power stayed on (or not), water stayed safe (or not), vulnerable customers were protected (or not). Counting policies, tabletop exercises, or modelled capacity *over-credits intentions*. We score events on the realized path and require *structural* closure of hazardous routes (HZ = 0) that is *proven* ($S^* = 1$).

25.2 Operational semantics (instantiation)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \,(\text{rights/safety/licence breach}) \text{ or } H_t \geq h_{\min}\big) \text{ and } C_t = 1.$$

$C_t = 1$ when an accountable operator/contractor/vendor had adequate knowledge & freedom and crossed a fault threshold (intent, recklessness, gross negligence) against a published *safety case*. If $C_t = 0$, record H_t for learning but do not set $R_t = 1$.

Harm H_t. Use domain scales: SAIDI/SAIFI-converted detriment[1], medically dependent outages (oxygen, dialysis), water-quality exceedances (E. coli, turbidity, chemicals), explosion/fire injury, sewage exposure, verified economic loss for small businesses, unlawful disconnections.

Mercy/Justice M_t, J_t^v. $M_t = 1$ when *delivered* restorations occur (priority reconnection, bottled water/home delivery, generators, compensation/credit, repairs to damaged equipment). $J_t^v = 1$ when rectification/accountability is *enacted* (protection lists fixed, switching logic/policies changed, contractor sanctions, regulator notifications).

Near-misses and drills. Near-miss (e.g., breaker trip averted by redundancy; contaminant alert caught pre-network) is not $R_t = 1$.

[1] System Average Interruption Duration/Interruption Frequency indices.

Convert into *eventized drills* ($D_t = 1$): black-start tests; network-islanding; contamination drills; cyber red-team exercises. Drills support S^* but do not mint M, J^v.

25.3 Evaluator and constraints (utilities view; broken to fit)

$$\mathcal{J}_{\text{util}} = \sum_{t \geq 0} \Big(\alpha\, \Delta L_t + \gamma\, \Delta F_t + \mu\, M_t + \nu\, J_t^v \Big) - \sum_{t \geq 0} \Big(\beta\, H_t + \kappa\, R_t \Big), \tag{25.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau \leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^* \in \{0, 1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t credits *delivered* protections (priority services, backup supply, rapid repair times, safe discharge); ΔF_t credits *exercised* rights (lawful connection, disconnection protections, timely information access).

25.4 Minimal-trigger doctrine (operator policy)

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (asset, protection list, control logic, contractor, comms), fix, tests,

owner+deadline, publication plan (privacy-preserving).

2. **Remedy:** Enact M_t (restoration/compensation, priority support, safe alternative supply) and J_t^v (rectification/accountability: policy/tool change, sanctions, regulator notice).

3. **Confirm:** Stronger-incentive reenactment: cohort replay → seasonal/weather/cyber stress → supervised live drills. On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats of c add cost (κ and H_t) but no new kind-level credit (typal/concave). One-off + closure dominates (Ch. 11).

25.5 Incident classes & scope locks (examples)

- **GRID-OUT** — unplanned outage/blackout beyond standards. *Scope:* protection schemes, vegetation, weather resilience, spares, call-out.
- **LOADSHED** — mis-prioritised load shedding harms protected sites (hospitals, critical services, medically dependent customers). *Scope:* lists, mapping, sequencing logic, comms.
- **GAS-LEAK** — leak/explosion hazard. *Scope:* detection, isolation, purging, response times, contractor competence.
- **WATER-QUAL** — water contamination/boil notice. *Scope:* treatment barriers, monitoring, network hydraulics, comms.

- **SEWER-OF** — sewage overflow/environmental discharge. *Scope:* pumping, capacity, wet-weather plans, alarms.

- **CYBER-ICS** — ICS/SCADA compromise impacting safety/reliability. *Scope:* network segmentation, MFA, allow-lists, manual fallback.

- **WORK-SAFE** — worker/public safety incident (electrical contact, confined space, trench collapse). *Scope:* permits, PPE, supervision, isolation.

- **BILL-ERR** — billing error/unlawful disconnection of protected/vulnerable customer. *Scope:* data quality, vendor interfaces, hold gates.

Each class requires a written *scope lock* freezing assets, thresholds, tools, and preconditions; new causal vectors open a new class c'.

25.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Injury/health impacts; outage duration counts (incl. protected sites); water-quality exceedances; environmental discharge volumes; verified economic loss; unlawful disconnection days
B_t (breach)	Licence/standard noncompliance; regulator orders; quality limit exceedance; black-start/load-shed plan violation
C_t (culpability)	Safety-case/process gap vs. standard; foreseeability; asset logs; change-control minutes; contractor oversight
M_t (mercy)	Priority reconnection; generators/water delivered; compensation/credits; repairs/replacements to damaged equipment
J_t^v (justice)	Policy/control-logic change; updated protection lists; sanctions; regulator notification; contractor remedies
ΔL_t	Achieved restoration times; backup provisioning; safe discharge operations; resilience works completed
ΔF_t	Rights exercised: reconnection, protected status honoured, access to information/appeals; disconnection holds applied
S_c^*	Passed replay + seasonal/cyber stress + supervised live drills; artefacts linked (privacy-preserving)

25.7 Confirmation tests (design pattern for utilities)

Step 1: Cohort replay. Re-run matched outages/incidents (same feeders/zones/plants) showing removal of the prior vector; include protected-site coverage.
Step 2: Stress batteries. Weather/capacity stress (storm, heat, freeze), supply constraints, cyber red-team on ICS with safe sandboxes; pre-registered pass metrics.
Step 3: Supervised live drills. Islanding/black-start drills, contamination response simulations with public-health observers, controlled load-shed exercises. On pass, set $S^*(c) = 1$.
Scope lock. Freeze class scope; later failures with different causality open c'.

25.8 Tool-gating and interim safeguards after a trigger

Until $S^*(c) = 1$: enforce disconnection holds for protected/vulnerable customers; install temporary generation/water bowsers; reduce transfer limits and exploration on auto-reconfiguration; require senior sign-off for switch-plan changes; increase inspection/vegetation patrols; enable manual fallback for ICS. Remove gates only after closure.

25.9 Micro-vignettes (worked examples)

(A) Drinking-water contamination (WATER-QUAL). *Trigger:* E. coli exceedance confirmed; boil notice issued; culpable process gap in treatment/monitoring $\Rightarrow R_t = 1$.
Remedy: Alternative supply delivered; compensation/credit; process/tool rectification; staff/accountability $(M_t, J_t^y = 1)$.
Confirm: Replay similar network states; seasonal stress with heavy rain; supervised drills with sampling; pass $\Rightarrow S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Mis-prioritised load shedding (LOADSHED). *Trigger:* Hospital and dialysis patients lose power due to stale protection list; rights breach and harm $\Rightarrow R_t = 1$.
Remedy: Priority reconnection; on-site generators; list governance re-built; sequencing logic revised; sanctions as needed $(M_t, J_t^y = 1)$.
Confirm: Cohort replay of feeder plans; black-start/load-shed drill with observers; closure on pass.

(C) SCADA compromise (CYBER-ICS). *Trigger:* Malware pivots to control network; unsafe setpoints sent; protection trips avert catastrophe (near-miss) then small spill causes harm $\Rightarrow R_t = 1$.
Remedy: Isolation/clean; allow-listing; MFA/segmentation; vendor contract sanctions; customer remediation $(M_t, J_t^y = 1)$.
Confirm: Red-team reenactment in cyber range; failover to manual; supervised live proving; class closed on pass.

25.10 Dashboards and metrics (ledger view)

- **Harm/recurrence:** SAIDI/SAIFI; medically dependent outage counts; water-quality noncompliance days; overflow volumes; recurrence by class; time-to-remedy; T_c^*.
- **Closure:** # classes with $S^* = 1$; median T_c^*; age/backlog of open dockets.
- **Culture (love):** priority support delivered; resilience works completed; customer comms performance (ΔL_t proxies).
- **Freedom:** protected disconnection holds; appeals upheld; access to information; reconnection SLAs met (ΔF_t).

25.11 Anti-gaming and integrity

- **Suppression control.** Mandatory incident registers; whistleblower channels; random field audits; environmental sampling; penalties for under-reporting.
- **Scope creep.** Lock class definitions; do not relabel repeats as "new" unless causal vector differs.
- **Vendor accountability.** Confirmation packs must include vendor artefacts (firmware, configs, test logs); no black-box exemptions.
- **Equity & vulnerability.** Require parity analysis for vulnerable/remote communities; confirm priority lists actually work in drills.

25.12 One-page checklist (drop-in for utilities)

Utility Incident → Confirmation Checklist

Trigger captured? assets/feeders/plants, timestamps, telemetry, samples

Gate set? protected-customer holds; temporary supply; reduced auto-reconfiguration; manual fallback enabled

Docket opened? root cause, fix, tests, owner & deadline, publish plan

Remedy enacted? restoration/compensation; priority services; accountability actions

Tests passed? cohort replay ✓ stress batteries (weather/cyber/capacity) ✓ supervised live drills ✓

Scope locked? class frozen; distinct vectors filed as c'

Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined

Gates removed? only post-closure; rationale logged

25.13 Limits and open problems

- **Cascades and interdependence.** Power–water–telecom coupling complicates C_t and closure; define cross-sector classes and joint drills.

- **Model error.** Forecasts of demand/hydraulics may drift; pair model updates with event-ledger audits.

- **Rare catastrophes.** Low-frequency high-impact events require heavy use of drills while retaining honesty locks (no M, J^v for drills).

25.14 What this chapter contributes

A full translation of the minimal-trigger doctrine to energy & utilities: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and evidence standards. With typal/concave credits and $\kappa > 0$, repeating the same vector adds cost without adding kind-level value; one-off and confirmed closure dominates.

Chapter 26

Environment & Climate: Discharges, Air, and Habitat Restoration

Notation

Domain mapping (environmental protection: rivers/coasts, air, soil, habitats, waste, climate).
An *event* is a verifiable environmental occurrence with adjudication potential (permit exceedance, discharge, spill, exceedance day, habitat loss, wildfire ignition/escape, illegal waste activity). $R_t \in \{0,1\}$: **rejection** (culpable rights/safety/permit breach or harm $\geq h_{\min}$); $B_t \in \{0,1\}$: breach (permit/standard/ESG/impact license); $C_t \in \{0,1\}$: culpability gate (intent/reckless/gross negligence by operator/authority/vendor); $H_t \geq 0$: realized harm

(biophysical harm or protected-rights harm: pollutant load, exceedance days, fish kill, habitat area lost, exposure of communities).
$M_t, J_t^y \in \{0,1\}$: *enacted* mercy/justice (cleanup, restoration, compensation; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: increments to care/stewardship (love) and exercised rights/voice (freedom).
Incident classes $c \in C$ (illustrative): **SEWER-OF** sewage/CSO discharge beyond permit; **IND-SPILL** industrial chemical spill; **BATH-NON** bathing water non-compliance; **AIR-PM** $PM_{2.5}/NO_2$ exceedance; **FISH-KILL** oxygen crash/toxicity kill; **HABITAT-LOSS** unlawful felling/land clearance; **WASTE-ILLEGAL** illegal dumping/burning; **WILDFIRE** preventable ignition/escape; **FLOOD-DEF** failure of maintained defences; **GHG-LEAK** venting/fugitive emissions beyond consent.
Switches: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0,1\}$ per class. Evaluator $\mathcal{J}$ is event-valued (Ch. 9).

26.1 Why an event-valued ledger fits environment & climate

Environmental assurance must answer *what actually happened*: pollutant mass entered water/air, habitat was lost or restored, exceedance days occurred or were prevented, communities were protected or harmed. Counting modelling capacity or policy intent over-credits plans. The ledger scores realized events, enacted remediation, and

proven structural closure of causal routes (HZ = 0, $S^* = 1$).

26.2 Operational semantics (instantiation)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \,(\text{permit/standard breach}) \,\text{or}\, H_t \geq h_{\min}\big) \,\text{and}\, C_t = 1.$$

$C_t = 1$ where an accountable operator/authority/vendor had adequate knowledge/freedom and crossed a fault threshold (intent, recklessness, gross negligence) relative to a published *safety/environmental case*. If $C_t = 0$, record H_t for learning but do not set $R_t = 1$.

Harm H_t. Choose transparent scales: pollutant mass (kg/day), exceedance days, dissolved oxygen minima, fish mortality counts/biomass, habitat area/condition units lost, population exposure counts, verified health/economic losses. Publish $h_{\min}$.

Mercy/Justice M_t, J_t^v (event-valued only). $M_t = 1$ when restoration is *delivered*: cleanup completed, habitat restored, compensation/alternative supply provided, community support. $J_t^v = 1$ when rectification/accountability is *enacted*: infrastructure upgrades, process/policy change, sanctions, regulator notices.

Near-misses and drills. Near-miss (e.g., alarm catches contaminant pre-release) is not $R_t = 1$; create *eventized drills* ($D_t = 1$): tracer tests, controlled valve simulations, forecast-based stress runs, wildfire exercises. Drills help S^* but do not mint M, J^v.

26.3 Evaluator and constraints (environmental view; broken to fit)

$$\mathcal{J}_{\text{env}} = \sum_{t\geq 0}\Big(\alpha\,\Delta L_t + \gamma\,\Delta F_t + \mu\,M_t + \nu\,J_t^{v}\Big) - \sum_{t\geq 0}\Big(\beta\,H_t + \kappa\,R_t\Big), \tag{26.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau\leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^{*} \in \{0,1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t credits *delivered* stewardship (km of river restored, barriers installed, green infrastructure built, public warnings delivered); ΔF_t credits *exercised* rights/voice (access to information, successful appeals, community monitoring access).

26.4 Minimal-trigger doctrine (operator/regulator policy)

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (asset capacity, control logic, maintenance, contractor, forecasting, enforcement), fix, tests, owner+deadline, publication plan (privacy-preserving).

2. **Remedy:** Enact M_t (cleanup, alternative supply, habitat repair, compensation) and J_t^{v} (rectifica-

tion/accountability: infrastructure upgrade, operating rules, sanctions, notices).

3. **Confirm:** Reenactment under stronger incentives: cohort replay → seasonal/weather stress → independent sampling or remote-sensing validation. On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats of c add cost (κ and H_t) but no new kind-level credit (typal/concave). One-off + closure dominates (Ch. 11).

26.5 Incident classes & scope locks (examples)

- **SEWER-OF** — sewage/CSO discharge beyond permit. *Scope:* storage, screens, storm capacity, telemetry, rainfall thresholds, duty cycles.

- **IND-SPILL** — industrial chemical spill to water/soil. *Scope:* bunding, valves, alarms, SOPs, spill kits, contractor controls.

- **BATH-NON** — bathing water non-compliance. *Scope:* pathogen indicators, sampling cadence, warning signage, upstream/downstream attribution.

- **AIR-PM** — $PM_{2.5}$/NO_2 exceedance days. *Scope:* emissions inventory, monitoring siting, traffic/industry episodes, episode response.

- **FISH-KILL** — oxygen crash/toxicity kill. *Scope:* abstraction rules, temperature flows, toxicity alarms, aeration deploy.

- **HABITAT-LOSS** — unlawful felling/clearance; net-gain commitments breached. *Scope:* permits, mapping, seasons, buffers, offsets.

- **WASTE-ILLEGAL** — illegal dumping/burning. *Scope:* surveillance, tracking, site security, seizure powers, vendor contracts.

- **WILDFIRE** — preventable ignition/escape from managed lands. *Scope:* burn plans, weather windows, fuel breaks, crew levels.

- **FLOOD-DEF** — maintained defence failure. *Scope:* inspection cadence, asset condition, emergency plans, spares.

- **GHG-LEAK** — fugitive/venting emissions beyond consent (landfill, oil/gas, refrigerants). *Scope:* LDAR, flare reliability, containment.

Each class has a written *scope lock* (assets, thresholds, tools, preconditions); new causal vectors open c'.

26.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Lab-confirmed pollutant loads; exceedance days; dissolved-oxygen minima; fish mortality counts; habitat area/condition loss; verified health/economic impact
B_t (breach)	Permit exceedance certificates; enforcement notices; court findings; independent audits; remote-sensing detections
C_t (culpability)	Safety-case/process gap vs. standard; foreseeability; maintenance/telemetry logs; change-control minutes
M_t (mercy)	Cleanup completed; alternative supply delivered; habitat restoration works started/completed; compensation paid
J_t^v (justice)	Infrastructure upgrade commissioned; control-logic/spec change; sanctions/fines; regulator notifications
ΔL_t	Stewardship delivered: km of river restored; wetland creation; tree canopy added; public warnings/comms actually issued
ΔF_t	Rights/voice exercised: access to data; upheld appeals; community monitoring portals; participation in hearings
S_c^*	Passed cohort replay + seasonal/weather stress + independent sampling/remote sensing (artefacts

26.7 Confirmation tests (design pattern)

Step 1: Cohort replay. Re-run matched assets/reaches/episodes (similar rain/flow/traffic/wind) showing removal of the prior vector.
Step 2: Seasonal/weather stress. Test in worst-season windows (first flush, heat waves, cold snaps, storm surges) or simulate with controlled flows/loads; include contractor/telemetry failure drills.
Step 3: Independent sampling/remote sensing. Third-party lab sampling; continuous sensors; satellite/airborne data; pre-registered pass metrics and thresholds.
Scope lock. Freeze class scope; different causal vectors open c'.

26.8 Tool-gating and interim safeguards after a trigger

Until $S^*(c) = 1$: throttle or suspend discharges; install temporary storage/aeration; deploy booms and absorbents; issue boil/no-swim notices; restrict operations to conservative windows; enable community sampling portals; require senior sign-off for risky operations. Remove gates only after closure.

26.9 Micro-vignettes (worked examples)

(A) CSO discharge beyond permit (SEWER-OF). *Trigger:* Storm overflow exceeds permit hours; pathogen loads high; culpable capacity/control gap $\Rightarrow R_t = 1$.
Remedy: Public warnings and alternative access; storage/monitoring

upgrade approved; compensation to affected businesses; river cleanup ($M_t, J_t^v = 1$).
Confirm: Cohort replay across storms; seasonal stress (first flush); independent sampling confirms compliance; $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Industrial solvent spill (IND-SPILL). *Trigger:* Bund failure; solvent to river; fish kill recorded; enforcement confirms fault $\Rightarrow R_t = 1$.
Remedy: Containment/cleanup; habitat restoration; sanctions; vendor controls ($M_t, J_t^v = 1$).
Confirm: Tracer tests; seasonal low-flow stress; third-party sampling; class closed on pass.

(C) $PM_{2.5}$ exceedances near schools (AIR-PM). *Trigger:* Monitors exceed limits on multiple days; episode plans not enacted; policy gap confirmed $\Rightarrow R_t = 1$.
Remedy: Episode-response rules enforced; traffic gating; retrofit programmes; school-route protections ($M_t, J_t^v = 1$).
Confirm: Matched-episode replay; meteorological stress; independent sensor network verification; closure on pass.

26.10 Dashboards and metrics (ledger view)

- **Harm/recurrence:** pollutant loads released; exceedance days; fish kills; habitat hectares lost; recurrence by class; time-to-remedy; time-to-closure T_c^*.

- **Closure:** # classes with $S^* = 1$; median T_c^*; age/backlog of open dockets.

- **Culture (love):** stewardship delivered (river km restored; trees planted; wetlands created); public warnings issued; community engagement (ΔL_t).

- **Freedom:** access-to-information fulfilled; appeals upheld; citizen-sensing portals live (ΔF_t).

26.11 Anti-gaming and integrity

- **Sampling games.** Require third-party labs, continuous sensors, and randomised timing; preserve raw telemetry with hashes.

- **Weather window cherry-pick.** Confirmation must include worst-season stress; document nuisance parameters.

- **Scope creep.** Lock class definitions; do not relabel repeats as "new" unless causal vector differs.

- **Transparency.** Publish privacy-preserving artefacts or allow independent audits; include remote-sensing where field access is limited.

- **Equity.** Track exposure and remediation by community (esp. vulnerable populations) in confirmation packs.

26.12 One-page checklist (drop-in for operators/regulators)

Environment Incident → Confirmation Checklist

Trigger captured? asset/reach IDs, timestamps, telemetry/samples, remote-sensing evidence

Gate set? throttles/suspensions; temporary storage/booms/aeration; public notices; conservative operating windows

Docket opened? root cause, fix, tests, owner & deadline, publish plan

Remedy enacted? cleanup; habitat restoration; alternative supply; compensation; accountability

Tests passed? cohort replay ✓ seasonal/weather stress ✓ third-party sampling/remote sensing ✓

Scope locked? class frozen; distinct vectors filed as c'

Closure set? $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$; monitoring window defined

Gates removed? only post-closure; rationale logged

26.13 Limits and open problems

- **Attribution.** Multiple sources and lagged transport complicate C_t; use mixing models and tracer evidence in confirmation.
- **Cumulative impacts.** Single-event closure may not redress

cumulative harm; maintain class families and basin/air-shed ledgers.

- **Rare catastrophes.** Low-frequency high-impact events force reliance on drills and remote sensing; preserve honesty locks (no M, J^v for drills).
- **Climate variability.** Confounding weather requires matched-episode methodology and robust nuisance-parameter logging.

26.14 What this chapter contributes

A full translation of the minimal-trigger doctrine to environment & climate: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and evidence standards. With typal/concave credits and $\kappa > 0$, repeating the same vector adds cost without adding kind-level value; one-off and confirmed closure dominates.

Chapter 27

Transport & Road Safety: Collisions, Designs, and Confirmation

Notation

Domain mapping (highways authorities, local roads, national networks, public transport, fleets, roadworks).
An *event* is a transport outcome with adjudication potential: collision, serious roadwork breach, unsafe speed environment, defect, crossing failure, unsafe PT operation, or data/privacy breach from transport tech.
$R_t \in \{0, 1\}$: **rejection** (culpable rights/safety breach or harm $\geq h_{\min}$); $B_t \in \{0, 1\}$: breach (standards/permit/code noncompliance: signage, taper, speed management, duty of care, operating

rules); $C_t \in \{0,1\}$: culpability gate (intent/recklessness/gross negligence by road authority/operator/contractor/fleet); $H_t \geq 0$: realized harm (KSI: killed/seriously injured; hospitalisation; verified economic loss; privacy harm).
$M_t, J_t^{\nu} \in \{0,1\}$: *enacted* mercy/justice (restoration/compensation and rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: increments to care/safety (love: protections installed) and exercised freedoms (freedom: safe access, mobility, due process).
Incident classes $c \in C$ (illustrative): **JUNC-KSI** high-harm junction vector; **SPEED-ENV** speed environment mis-set; **VRU-CROSS** pedestrian/cycle crossing failures; **WRK-ZONE** work-zone intrusion/protection failure; **FLEET-FAT** professional driver fatigue/roster gap; **VEH-DEF** unsafe vehicle maintenance; **PT-SAFETY** public-transport door/boarding/platform gap; **SCHOOL-ZONE** inadequate school-street protections; **WINTER-MAINT** ice/grit failure; **DATA-PRIV** ANPR/telematics privacy breach.
Switches: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0,1\}$ per class. Evaluator $\mathcal{J}$ is event-valued (Ch. 9).

27.1 Why an event-valued ledger fits transport

Road safety is judged by *what happened*: collisions prevented, hazards removed, and routes *proven* closed. Counting policies, awareness campaigns, or modelled capacity over-credits intentions.

The ledger scores realized events (H_t, R_t), enacted remedies (M_t, J_t^v), and structural closure (HZ = 0, $S^* = 1$).

27.2 Operational semantics (instantiation)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \text{ (standards/duty breach) or } H_t \geq h_{\min}\big) \text{ and } C_t = 1.$$

$C_t = 1$ when an accountable road owner/operator/contractor/fleet had adequate knowledge & freedom and crossed a fault threshold (intent, recklessness, gross negligence). If $C_t = 0$ (accident), log H_t for learning but do not set $R_t = 1$.

Harm H_t. Use transparent scales: KSI counts, hospitalisations, property damage above threshold, verified economic loss, access deprivation days (e.g., severed crossings), privacy harms. Publish $h_{\min}$.

Mercy/Justice M_t, J_t^v. $M_t = 1$ when restoration is *delivered* (compensation, medical/rehab support, access restored, apology). $J_t^v = 1$ when *rectification/accountability* is enacted (design/tool/process change, sanctions, regulator notices).

Near-misses and drills. Near-miss (conflicts, red-light run stopped by margin, work-zone incursion averted) is not $R_t = 1$; convert into *eventized drills* ($D_t = 1$): controlled speed trials, work-zone barrier tests, conflict-analysis drills. Drills support S^* but do not mint M, J^v.

27.3 Evaluator and constraints (transport view; broken to fit)

$$\mathcal{J}_{\text{trs}} = \sum_{t\geq 0}\Big(\alpha\,\Delta L_t + \gamma\,\Delta F_t + \mu\,M_t + \nu\,J_t^{v}\Big) - \sum_{t\geq 0}\Big(\beta\,H_t + \kappa\,R_t\Big), \tag{27.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau\leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^{*} \in \{0,1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t credits *delivered* protections (geometry fixes, crossings installed, barriers, signage, gritting done); ΔF_t credits *exercised* freedoms (safe crossing access, bus-stop access, appeal outcomes, transparent data rights).

27.4 Minimal-trigger doctrine (authority/operator policy)

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (geometry, speed setting, protection, roster, maintenance, comms), fix, tests, owner+deadline, publication plan.

2. **Remedy:** Enact M_t (restoration/compensation, access restored) and J_t^{v} (rectification/accountability: design change, tool-gates, sanctions, notices).

3. **Confirm:** Stronger-incentive reenactment: matched-site replay → seasonal/weather/peak stress → monitored live with conflict analytics. On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats of c add cost (κ and H_t) but no new kind-level credit (typal/concave); one-off + closure dominates (Ch. 11).

27.5 Incident classes & scope locks (examples)

- **JUNC-KSI** — repeated KSI at a junction/roundabout. *Scope:* visibility, approach speeds, turn conflicts, signal phases, protection for VRUs.

- **SPEED-ENV** — mis-set speed environment (posted/design speed misaligned; 85th percentile excessive). *Scope:* geometry, frontage, crossings, enforcement.

- **VRU-CROSS** — pedestrian/cycle crossing failure. *Scope:* desire lines, wait times, refuge/island, lighting, detection.

- **WRK-ZONE** — work-zone intrusion/protection failure. *Scope:* tapers, barriers, buffer lengths, speed limits, lookouts.

- **FLEET-FAT** — fatigue/roster failure in fleets (HGV/coach/bus). *Scope:* hours, scheduling, telematics, monitoring, routes.

- **VEH-DEF** — vehicle defect/maintenance failure. *Scope:*

inspection cadence, brake/tyre records, defect response, contractor.

- **PT-SAFETY** — public-transport stop/door/platform gap hazards. *Scope:* platform geometry, dispatch rules, door interlocks, crowding.

- **SCHOOL-ZONE** — school-street protections missing/weak. *Scope:* timings, access controls, crossings, parking controls, marshals.

- **WINTER-MAINT** — ice/grit failure on known risk segments. *Scope:* route priority, sensors, salt stocks, call-out triggers.

- **DATA-PRIV** — ANPR/telematics privacy/security breach. *Scope:* DPIA, retention, access controls, vendor links.

Each class must include a written *scope lock*; new causal vectors open c' rather than relabeling.

27.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Police collision reports; hospital data; KSI counts; verified economic loss; service disruption minutes; access deprivation logs
B_t (breach)	Standards/code noncompliance; audit findings; regulator/inspector outcomes; device/roster/maintenance logs
C_t (culpability)	Safety-case gap vs. standard; foreseeability; historic complaints; speed distributions; design drawings/records
M_t (mercy)	Compensation/restoration; rehab support; access reinstated (crossings/bus stops); apology/record correction
J_t^v (justice)	Design/process/tool change shipped; enforcement/roster fix; sanctions; regulator notices
ΔL_t	Protections delivered: crossings, refuge, barriers, signals/phasing, grit cycles; work-zone protections installed
ΔF_t	Rights exercised: access restored, appeals upheld, privacy/data rights fulfilled
S_c^*	Passed matched-site replay + seasonal/peak stress + monitored live with conflict analytics/speed checks

27.7 Confirmation tests (design pattern for roads/operators)

Step 1: Matched-site replay. Compare before/after at the site and against matched controls (similar AADT, mix, geometry) to show prior vector removed.
Step 2: Seasonal/peak stress. Test under dark/wet/ice and peak-hour volumes; include special events; verify protection holds.
Step 3: Monitored live. Short live window with independent observers, speed surveys, conflict analytics (e.g., post-encroachment time, harsh braking); pre-registered pass metrics.
Scope lock. Freeze class scope; new causal vectors open c'.

27.8 Tool-gating and interim safeguards after a trigger

Until $S^*(c) = 1$: temporary speed limits; portable traffic-calming; temporary signals; marshals; interim barriers/tapers; night-only works; fleet route restrictions and rest gates; stop-gap bus-boarding controls. Remove gates only after closure.

27.9 Micro-vignettes (worked examples)

(A) Junction with repeated KSI (JUNC-KSI). *Trigger:* Multiple serious collisions, turning conflicts; audit shows visibility/speed mismatch $\Rightarrow R_t = 1$.
Remedy: Geometry fix (tighten radii), signal phasing for protected

turns, refuge islands, interim speed reduction ($M_t, J_t^v = 1$).
Confirm: Matched-site replay; peak/dark stress; monitored live with speed/conflict analytics; $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Work-zone intrusion (WRK-ZONE). *Trigger:* Vehicle enters work-zone, injuring worker; taper/barrier noncompliant $\Rightarrow B_t = 1$, $C_t = 1$, so $R_t = 1$.
Remedy: Compensation; barrier/taper spec upgrade; speed gate; lookout rule; contractor sanctions ($M_t, J_t^v = 1$).
Confirm: Night/peak drills; monitored live with intrusion sensors; class closed on pass.

(C) School-street speeding (SCHOOL-ZONE). *Trigger:* Child injury near school; missing access controls; process gap confirmed $\Rightarrow R_t = 1$.
Remedy: Timed access closure, raised table/crossings, marshal programme, enforcement ($M_t, J_t^v = 1$).
Confirm: Term-time peak tests; speed surveys; random patrols; closure on pass.

27.10 Dashboards and metrics (ledger view)

- **Harm/recurrence:** KSI/serious collisions; recurrence by class; time-to-remedy; time-to-closure T_c^*.
- **Closure:** # classes with $S^* = 1$; median T_c^*; backlog/age of open dockets.

- **Culture (love):** protections installed; gritting delivery; work-zone compliance; VRU upgrades (ΔL_t proxies).

- **Freedom:** access restored; appeals upheld; privacy/data rights fulfilled (ΔF_t).

27.11 Anti-gaming and integrity

- **Regression-to-mean.** Use matched controls and exposure metrics (volume, composition) in confirmation; pre-register pass criteria.

- **Under-reporting.** Pair police reports with hospital/insurance/community data; require random audits.

- **Speed laundering.** Publish full distributions, not means; use independent surveys.

- **Scope creep.** Lock class definitions; do not relabel repeats as "new" unless causal vector differs.

- **Privacy.** For ANPR/telematics, publish artefacts with redactions or enable independent audit.

27.12 One-page checklist (drop-in for roads/operators)

Transport Incident → Confirmation Checklist

Trigger captured? location, time, geometry/roster/maintenance logs, speed/conflict data
Gate set? temporary speed/calming; barriers/tapers; marshals; night-only works; fleet route/rest gates
Docket opened? root cause, fix, tests, owner & deadline, publish plan
Remedy enacted? compensation/restoration; design/process change; accountability
Tests passed? matched replay ✓ seasonal/peak stress ✓ monitored live with analytics ✓
Scope locked? class frozen; distinct vectors filed as c'
Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined
Gates removed? only post-closure; rationale logged

27.13 Limits and open problems

- **Attribution.** Multi-actor causality (authority, contractor, fleet, enforcement) complicates C_t; adopt shared-fault taxonomies.
- **Exposure drift.** Volume/mix changes confound comparisons; normalise by exposure in confirmation.

- **Rare but severe events.** Low-frequency catastrophes require heavy use of drills while retaining honesty locks (no M, J^{v} for drills).

27.14 What this chapter contributes

A full translation of the minimal-trigger doctrine to transport & road safety: precise triggers (R_t), docketed remedies (M, J^{v}), structural closure (S^{*}), and evidence standards. With typal/concave credits and $\kappa > 0$, repeating the same vector adds cost without adding kind-level value; one-off and confirmed closure dominates. “‘□0■

Chapter 28

Digital Platforms & Online Safety: Violations, Red-Teaming, and Confirmation

Notation

Domain mapping (consumer platforms, social media, messaging, live-streaming, app stores, marketplaces).
An *event* is a platform outcome with adjudication potential (policy violation, harmful recommendation, unlawful processing, fraud, safety incident).
$R_t \in \{0, 1\}$: **rejection** (culpable rights/safety breach or harm $\geq h_{\min}$); $B_t \in \{0, 1\}$: breach (terms/standard/law: child-

safety, illegal content, fraud, privacy, algorithmic transparency); $C_t \in \{0,1\}$: culpability gate (intent/recklessness/gross negligence by the operator/vendor/mod contractor); $H_t \geq 0$: realized harm (child-safety harm, financial loss, harassment/abuse harm, privacy harm, verified medical/physical harm).
$M_t, J_t^y \in \{0,1\}$: *enacted* mercy/justice (restoration to the actually harmed; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: love/freedom increments (care delivered; rights exercised in practice).
Incident classes $c \in C$ (illustrative): **CHILD-SAFE** child-safety failure (grooming/unsafe contact signals); **SELF-HARM** pro-self-harm content promoted; **EXTREM-AMP** violent/extremist promotion; **DOX-PRIV** privacy/data breach (doxing, unlawful processing); **FRAUD-SCAM** consumer fraud/impersonation; **REC-FAIL** recommender amplification of prohibited content; **ADS-FAIL** ad policy to minors/illegal targeting; **LIVE-MOD** live-stream moderation failure; **MARKET-RISK** unsafe/illegal goods listings; **BOT-NET** coordinated inauthentic behaviour.
Switches: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0,1\}$ per class. Evaluator $\mathcal{J}$ is event-valued (Ch. 9).

28.1 Why an event-valued ledger fits platforms

Platform safety is often reported as *policy capacity* (filters, classifiers, T&C) rather than what users actually experienced. Our ledger counts realized violations and harms, enacted remedies, and whether

hazardous vectors are *structurally* and *provably* closed (HZ = 0, $S^* = 1$). Near-miss detection, simulations, and red-teaming help confirm fixes but do not mint mercy/justice credit.

28.2 Operational semantics (instantiation)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \,(\text{policy/legal breach}) \,\text{or}\, H_t \geq h_{\min}\big) \,\text{and}\, C_t = 1.$$

$C_t = 1$ where an accountable operator (platform, vendor, moderator contractor, ad broker) had adequate knowledge/freedom and crossed a fault threshold (intent, recklessness, gross negligence) relative to a published *safety case*. If $C_t = 0$, record H_t for learning but do not set $R_t = 1$.

Harm H_t. Use transparent scales: verified financial loss (fraud/impersonation), documented harassment/abuse outcomes, exposure hours for minors to harmful categories, privacy harms (doxing, unlawful data use), medical/physical harm attributable to platform action.

Mercy/Justice M_t, J_t^y. $M_t = 1$ when *delivered* restoration occurs (refund/compensation, takedown with user support, account restoration, privacy remediation). $J_t^y = 1$ when rectification/accountability is *enacted* (policy/tool change, sanctions, vendor termination, regulator notice).

Near-misses & drills. Near-miss (blocked upload, classifier caught attempt) is not $R_t = 1$; log and convert to *eventized drills* ($D_t = 1$): red-team prompts/scenarios, sandboxed uploads, shadow deployments. Drills support S^* but do not mint M, J^v.

28.3 Evaluator and constraints (platform view; broken to fit)

$$\mathcal{J}_{\text{plat}} = \sum_{t \geq 0} \Big(\alpha \, \Delta L_t + \gamma \, \Delta F_t + \mu \, M_t + \nu \, J_t^v \Big) - \sum_{t \geq 0} \Big(\beta \, H_t + \kappa \, R_t \Big), \tag{28.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau \leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^* \in \{0, 1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t credits *delivered* protections (age gates, friction, content limits, safety response time); ΔF_t credits *exercised* rights (appeals upheld, data access/erasure fulfilled, transparent notices).

28.4 Minimal-trigger doctrine (operator policy)

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (model, thresholds, feedback loops, vendor/tool, process, policy), fix, tests,

owner+deadline, publication plan.

2. **Remedy:** Enact M_t (refund/support/takedown/account repair) and J_t^v (rectification/accountability: model retrain, threshold gates, UI friction, vendor sanctions, regulator notice).

3. **Confirm:** Stronger-incentive reenactment: cohort replay → adversarial red-team → supervised live/shadow. On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats of c add cost (κ and H_t) but no new kind-level credit (typal/concave). One-off + closure dominates (Ch. 11).

28.5 Incident classes & scope locks (examples)

- **CHILD-SAFE** — failure to protect minors (unsafe contact/exposure). *Scope:* age/identity checks, messaging gates, content filters, reporting.

- **SELF-HARM** — recommend/boost pro-self-harm content. *Scope:* recommender signals, thresholds, safety responses, signposting.

- **EXTREM-AMP** — amplification of violent/extremist content. *Scope:* policy lists, classifier coverage, recency/virality caps.

- **DOX-PRIV** — doxing/unlawful processing or data leak. *Scope:* DPIA, access controls, vendor links, retention.

- **FRAUD-SCAM** — impersonation/fraud causing loss. *Scope:* identity verification, payments, chargeback, trust/safety.

- **REC-FAIL** — recommender boosts prohibited categories. *Scope:* objectives, feedback loops, downranking, holdback tests.

- **ADS-FAIL** — non-compliant ad targeting to minors/illegal categories. *Scope:* age gates, category bans, broker controls.

- **LIVE-MOD** — live-stream moderation failure. *Scope:* pre-triage, delay buffers, escalation, kill-switches, supervisor gates.

- **MARKET-RISK** — unsafe/illegal goods listings. *Scope:* seller gating, recalls, takedown SLAs, sampling.

- **BOT-NET** — coordinated inauthentic behaviour. *Scope:* graph signals, rate limits, verification, takedown cadence.

Each class has a written *scope lock* freezing inputs, thresholds, tools, and preconditions; new causal vectors open c' rather than relabeling.

28.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Verified user harm reports; financial loss proofs; exposure-hour metrics for minors; documented harassment outcomes; privacy harm records
B_t (breach)	Policy/legal breach findings; regulator/ombudsman decisions; audit logs; vendor/broker records
C_t (culpability)	Safety-case/process gap vs. standard; foreseeability; change-control minutes; classifier/threshold settings
M_t (mercy)	Refund/compensation; account restoration; takedown with user support; privacy remediation; apologies/notices
J_t^v (justice)	Model/policy/UI change shipped; sanctions/vendor terminations; regulator notices; transparency reports
ΔL_t	Safety frictions enabled; response-time SLAs met; age/identity checks working; help resources delivered
ΔF_t	Appeals upheld; data access/erasure fulfilled; clear notices; effective blocks/mutes provided
S_c^*	Passed cohort replay + red-team + supervised live/shadow; artefacts linked (privacy-preserving)

28.7 Confirmation tests (design pattern for platforms)

Step 1: Cohort replay. Re-run matched cohorts (same languages/locales/age buckets) showing removal of the prior vector; include holdout controls.
Step 2: Red-team/adversarial. Prompt/model attacks; coordinated upload tests; marketplace seeding; bot-net emulation; publish pass/fail artefacts.
Step 3: Supervised live/shadow. Short live window with enhanced supervision (kill-switches, rate caps, human-in-the-loop) and pre-registered pass metrics.
Scope lock. Freeze class scope; later failures with different causality open c'.

28.8 Tool-gating and interim safeguards after a trigger

Until $S^*(c) = 1$: throttle virality; raise thresholds; add friction/age gates; restrict live-streaming; pause risky ad categories; quarantine vendors; enable stronger identity checks; require senior sign-off for sensitive changes. Remove gates only after closure.

28.9 Micro-vignettes (worked examples)

(A) Recommender boosts prohibited content (REC-FAIL). *Trigger:* Audit finds feed boosting a prohibited class for minors;

exposure hours exceed threshold; policy gap confirmed $\Rightarrow R_t = 1$.
Remedy: Downranking and safety frictions; child-safety review; notices to affected users; transparency report ($M_t, J_t^v = 1$).
Confirm: Cohort replay; adversarial red-team; supervised live with kill-switch; $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Fraud/impersonation causing loss (FRAUD-SCAM). *Trigger:* Users suffer verified losses via impersonation accounts; KYC/verification gap confirmed $\Rightarrow R_t = 1$.
Remedy: Refunds/compensation; account restoration; verified-badge + sign-up friction; broker sanctions ($M_t, J_t^v = 1$).
Confirm: Replay with seeding; red-team identity attacks; shadow deployment under higher attack rates; closure on pass.

(C) Live moderation failure (LIVE-MOD). *Trigger:* Harmful live content persists beyond SLA; escalation failed; culpability found $\Rightarrow R_t = 1$.
Remedy: Delay buffer; human escalation; kill-switch policy; staffing floor; public notices ($M_t, J_t^v = 1$).
Confirm: Simulated surges; monitored live drills; class closed on pass.

28.10 Dashboards and metrics (ledger view)

- **Harm/recurrence:** verified harm reports; exposure-hour breaches; fraud losses; recurrence by class; time-to-remedy; T_c^*.

- **Closure:** # classes with $S^* = 1$; median T_c^*; backlog/age of open dockets.

- **Culture (love):** safety frictions enabled; response SLAs; transparency updates; user support delivered (ΔL_t proxies).

- **Freedom:** appeals upheld; data rights fulfilled; effective user controls (ΔF_t).

28.11 Anti-gaming and integrity

- **Suppression control.** Mandatory incident registers; protected disclosures; independent audits; penalties for under-reporting.

- **Metric laundering.** Publish distributions (exposure hours, takedown delay) not only means; pre-register pass criteria.

- **Scope creep.** Lock class definitions; do not relabel repeats as "new" unless causal vector differs.

- **Privacy.** Publish artefacts with redactions or enable independent audit when public release risks confidentiality.

28.12 One-page checklist (drop-in for platform operators)

Platform Incident → Confirmation Checklist

Trigger captured? cohort, locale, device/app versions, artefacts

Gate set? virality caps; frictions; age/identity gates; live-stream restrictions; ad-category pauses

Docket opened? root cause, fix, tests, owner & deadline, publish plan

Remedy enacted? refunds/compensation; takedown/support; account repair; accountability

Tests passed? cohort replay ✓ adversarial red-team ✓ supervised live/shadow ✓

Scope locked? class frozen; distinct vectors filed as c'

Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined

Gates removed? only post-closure; rationale logged

28.13 Limits and open problems

- **Attribution.** Multi-actor causality (platform, ad broker, vendor, user networks) complicates C_t; adopt shared-fault taxonomies.

- **Adversarial drift.** Attackers adapt; require periodic red-teams

and rolling confirmations for high-risk classes.

- **Measurement.** Hidden harms (psychological, social) require qualitative panels alongside quantitative KPIs.

28.14 What this chapter contributes

A full translation of the minimal-trigger doctrine to digital platforms & online safety: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and evidence standards. With typal/concave credits and $\kappa > 0$, repeating the same vector adds cost without adding kind-level value; one-off and confirmed closure dominates.

Chapter 29

Employment & Welfare (DWP): Decisions, Payments, and Confirmation

Notation

Domain mapping (employment & welfare systems: jobcentres, benefits administration, disability assessments, sanctions/appeals).

An *event* is a claimant-relevant outcome with adjudication potential (decision, payment, assessment, sanction, appeal, safeguarding, data handling).

$R_t \in \{0,1\}$: **rejection** (culpable rights/standards breach or

harm $\geq h_{\min}$); $B_t \in \{0,1\}$: breach (legal/policy/standard: unlawful decision, payment failure, appeal delay, accessibility failure, privacy breach); $C_t \in \{0,1\}$: culpability gate (intent/reckless/gross negligence by the authority/contractor/vendor); $H_t \geq 0$: realized harm (arrears, homelessness risk days, food insecurity days, debt collection, health impact, privacy harm).

$M_t, J_t^v \in \{0,1\}$: enacted mercy/justice (restoration to the *actually* harmed; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: increments to care/support (love) and exercised rights/freedoms (freedom).

Illustrative incident classes $c \in C$: **DEC-ERR** incorrect entitlement decision; **PAY-ERR** payment error/delay; **SANCT-WRONG** wrongful sanction; **ACCESS-FAIL** reasonable adjustment/communication failure; **APPEAL-DEL** appeal/mandatory reconsideration delay beyond standard; **ALGO-ERR** algorithmic decision error or profiling breach; **SAFEG-FAIL** safeguarding failure for vulnerable claimant; **DATA-PRIV** privacy/data breach; **RTW-GAP** return-to-work support gap causing detriment.

Switches: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0,1\}$ per class. Evaluator $\mathcal{J}$ is event-valued (Ch. 9).

29.1 Why an event-valued ledger fits employment & welfare

Welfare performance is often reported as *capacity* (service levels, headcount, digital portals). Claimants experience *events*: was the decision correct, was the payment on time, did the appeal resolve, were adjustments delivered? Our ledger scores realized outcomes (H_t, R_t), enacted remedies (M_t, J_t^v), and whether hazardous routes are *structurally* closed (HZ = 0) and *proven* closed ($S^* = 1$).

29.2 Operational semantics (instantiation)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \,(\text{rights/standard breach}) \,\text{or}\, H_t \geq h_{\min}\big) \,\text{and}\, C_t = 1.$$

$C_t = 1$ when an accountable actor (department/agency/contractor/vendor) had adequate knowledge/freedom and crossed a fault threshold (intent, recklessness, gross negligence). If $C_t = 0$ (accident), record H_t for learning but do not set $R_t = 1$.

Harm H_t. Transparent scales: days of arrears, missed rent/council tax, verified food bank reliance, debt collection/interest, loss of healthcare/transport access, homelessness nights, documented health impacts, privacy harms. Publish $h_{\min}$.

Mercy/Justice M_t, J_t^v. $M_t = 1$ when restoration is *delivered*: backdated payments with interest, hardship/short-term benefit advances,

debt/collection holds, record corrections, apology/compensation. $J_t^v = 1$ when *rectification/accountability* occurs: policy/process/tool change, sanctions/contract actions, regulator/ombudsman notices.

Near-misses and drills. Near-miss (auto-flagged payment anomaly fixed before harm) is not $R_t = 1$; convert to *eventized drills* ($D_t = 1$): sandbox reruns, red-team profiles, queue-load tests. Drills support S^* but do not mint M, J^v.

29.3 Evaluator and constraints (DWP view; broken to fit)

$$\mathcal{J}_{\text{dwp}} = \sum_{t\geq 0} \Big(\alpha\,\Delta L_t + \gamma\,\Delta F_t + \mu\,M_t + \nu\,J_t^v\Big) - \sum_{t\geq 0} \Big(\beta\,H_t + \kappa\,R_t\Big), \tag{29.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau\leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^* \in \{0,1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t credits *delivered* support (hardship payments, casework assistance, safeguarding); ΔF_t credits *exercised* rights (reasonable adjustments, appeals upheld, data/record rights).

29.4 Minimal-trigger doctrine (authority policy)

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (rules engine, assessment vendor, queueing, data match, communications, accessibility), fix, tests, owner+deadline, publication plan (privacy-preserving).

2. **Remedy:** Enact M_t (backpay with interest, hardship advance, debt/eviction hold, record correction, apology) and J_t^y (rectification/accountability: policy/tool change, sanctions/contract remedies, regulator/ombudsman notice).

3. **Confirm:** Stronger-incentive reenactment: cohort replay $\rightarrow$ stress red-team (edge cases, disability communication modes) $\rightarrow$ monitored live. On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats add cost (κ and H_t) but no new kind-level credit (typal/concave); one-off + closure dominates (Ch. 11).

29.5 Incident classes & scope locks (examples)

- **DEC-ERR** — incorrect decision (e.g., disability assessment error). *Scope:* assessor standards, evidence use, audit sampling, reversal rates.
- **PAY-ERR** — payment delay/error/underpayment. *Scope:* rules engine, bank integrations, cut-off logic, queueing.
- **SANCT-WRONG** — wrongful sanction (good-cause ignored). *Scope:* evidence gates, discretion policy, review cadence.
- **ACCESS-FAIL** — failure to provide reasonable adjustments/accessible communications. *Scope:* formats, interpreters, assisted digital, home visits.
- **APPEAL-DEL** — appeal/MR delay beyond standard causing harm. *Scope:* triage, prioritisation (risk-of-harm), scheduling.
- **ALGO-ERR** — algorithmic decision/profiling error or unlawful use. *Scope:* model governance, fairness checks, explanations, redress.
- **SAFEG-FAIL** — safeguarding failure (vulnerable claimants: domestic abuse, homelessness risk). *Scope:* flags, escalation, multi-agency coordination.
- **DATA-PRIV** — privacy breach/unlawful processing. *Scope:* DPIA, access controls, vendor links, retention.

- **RTW-GAP** — return-to-work support gap causing detriment. *Scope:* referral timing, training slots, employer links, conditionality balance.

Each class has a written *scope lock* (inputs, thresholds, tools, preconditions); new causal vectors open c'.

29.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Rent arrears/eviction proceedings; debt collection; missed essentials; homelessness nights; verified health impacts; privacy harms
B_t (breach)	Tribunal/ombudsman findings; audit fails; unlawful processing decisions; appeal timeliness breaches; accessibility failures
C_t (culpability)	Safety-case/process gap vs. standard; foreseeability; queue/decision logs; vendor contracts; governance minutes
M_t (mercy)	Backpay with interest; hardship supports; debt/eviction holds; record corrections; apology/compensation
J_t^v (justice)	Policy/tool/process change shipped; sanctions/contract remedies; regulator/ombudsman notices
ΔL_t	Delivered supports: advances, casework, safeguarding actions, proactive communication
ΔF_t	Rights exercised: adjustments provided, appeals upheld, data/record rights fulfilled
S_c^*	Passed cohort replay + stress red-team + monitored live; artefacts linked (privacy-preserving)

29.7 Confirmation tests (design pattern for welfare systems)

Step 1: Cohort replay. Re-run matched cases (same benefit types, demographics, disability profiles) showing removal of the prior vector; include historical controls.
Step 2: Stress red-team. Edge-case injections (irregular income, mixed household, disability communications); vendor swap tests; queue spikes; accessibility modes; pre-registered pass metrics.
Step 3: Monitored live. Short live window with enhanced supervision (debt holds by default, faster hardship routes, independent observers). On pass, set $S^*(c) = 1$.
Scope lock. Freeze class scope; later failures with different causality open c'.

29.8 Tool-gating and interim safeguards after a trigger

Until $S^*(c) = 1$: enable automatic debt/eviction holds, pay *on account* where lawful, fast-track hardship advances, add senior sign-off for sanctions, default to accessible communications, escalate safeguarding. Remove gates only after closure.

29.9 Micro-vignettes (worked examples)

(A) Payment delay causing arrears (PAY-ERR). *Trigger:* Rules engine cut-off misapplied; claimant misses rent; arrears escalate;

culpability confirmed $\Rightarrow R_t = 1$.
Remedy: Backpay with interest; hardship support; debt/eviction hold; apology ($M_t = 1$). Rectification ($J_t^v = 1$): cut-off logic fix, queue rules, vendor patch.
Confirm: Cohort replay around cut-off dates; stress spikes; monitored live period; $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Wrongful sanction (SANCT-WRONG). *Trigger:* Good-cause evidence ignored; sanction applied; food insecurity results; tribunal overturns $\Rightarrow R_t = 1$.
Remedy: Backpay; hardship; record correction; apology ($M_t = 1$); rectification ($J_t^v = 1$): discretion policy change, review gates, training/supervision.
Confirm: Red-team files with good-cause variants; monitored live with senior review; closure on pass.

(C) Algorithmic profiling error (ALGO-ERR). *Trigger:* Risk model wrongly flags fraud; payment paused; privacy/rights breached; governance finds model gap $\Rightarrow R_t = 1$.
Remedy: Payment restored; compensation; explanations provided; model retrain with fairness guardrails ($M_t, J_t^v = 1$).
Confirm: Shadow evaluation with protected groups; adversarial probes; supervised live; class closed on pass.

29.10 Dashboards and metrics (ledger view)

- **Harm/recurrence:** arrears days; homelessness nights; wrongful sanctions; decision-overturn rates; recurrence by class; time-to-remedy; time-to-closure T_c^*.
- **Closure:** # classes with $S^* = 1$; median T_c^*; age/backlog of open dockets.
- **Culture (love):** hardship supports delivered; safeguarding actions; proactive comms; casework load (ΔL_t proxies).
- **Freedom:** adjustments provided; appeals upheld; data/record rights fulfilled; timely access to work support (ΔF_t).

29.11 Anti-gaming and integrity

- **Under-reporting.** Mandatory incident registers; protected disclosures; random audits across vendors/offices.
- **Metric laundering.** Publish distributions (delays, overturn times) not only means; pre-register pass criteria.
- **Scope creep.** Lock class definitions; do not relabel repeats as "new" unless causal vector differs.
- **Equity.** Confirm closure holds across disability/health/protected groups; report subgroup metrics.

29.12 One-page checklist (drop-in for departments/agencies)

Welfare Incident → Confirmation Checklist

Trigger captured? benefit type, timestamps, decision/payment logs, artefacts

Gate set? debt/eviction holds; hardship advances; accessible communications; senior sign-off for sanctions

Docket opened? root cause, fix, tests, owner & deadline, publish plan

Remedy enacted? backpay/interest; hardship supports; record correction; accountability

Tests passed? cohort replay ✓ stress red-team ✓ monitored live ✓

Scope locked? class frozen; distinct vectors filed as c'

Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined

Gates removed? only post-closure; rationale logged

29.13 Limits and open problems

- **Attribution.** Multi-actor causality (department, contractor, bank/vendor, tribunal) complicates C_t; adopt shared-fault taxonomies.

- **Data quality.** Incomplete logs and manual overrides impede

evidence; invest in auditable capture and explainability for models.

- **Rare but severe harms.** Low-frequency catastrophic harms (eviction, health crises) require high-fidelity drills and priority gates while maintaining honesty locks.

29.14 What this chapter contributes

A full translation of the minimal-trigger doctrine to employment & welfare: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and evidence standards. With typal/concave credits and $\kappa > 0$, repeating the same vector adds cost without adding kind-level value; one-off and confirmed closure dominates.

Part V

Casebooks: Deep Dives

Chapter 30

Casebook: Justice & Courts: Disclosure, Delay, and Confirmation

Notation

Domain mapping (police → prosecution → courts → prisons/probation).
An *event* is a justice-system occurrence with adjudication potential (charging, disclosure, hearing, verdict, appeal, custody decision, data handling).
$R_t \in \{0,1\}$: **rejection** (culpable rights/standards breach or harm $\geq h_{\min}$); $B_t \in \{0,1\}$: breach (law/procedure/code: disclosure, timely trial, due process, equality of arms, safeguarding, privacy/data); $C_t \in \{0,1\}$: culpability gate (intent/reckless/gross

negligence by police/prosecution/courts/service); $H_t \geq 0$: realized harm (wrongful deprivation of liberty, unfair trial harm, victim harm from delay/collapse, privacy harms, verified financial loss).
$M_t, J_t^v \in \{0, 1\}$: *enacted* mercy/justice (restoration to the actually harmed; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: increments to care/justice delivered (love) and exercised rights/freedoms (freedom).
Incident classes $c \in C$ (illustrative): **DISC-FAIL** disclosure failure; **CASE-DEL** excessive case delay; **REMAND-EX** excessive/unlawful remand; **WCONV** wrongful conviction; **EVID-LOSS** evidence loss/spoliation; **INT-FAIL** interpreter/translation failure; **SAFEG-WIT** witness/victim safeguarding failure; **DATA-PRIV** privacy/data breach; **DIGI-EVID** digital-evidence ingestion/search failure; **CP-CHARGE** charging decision error.
Switches: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0, 1\}$ per class. Evaluator $\mathcal{J}$ is event-valued (Ch. 9).

30.1 Why an event-valued ledger fits justice

Justice is not a stack of policy binders; it is whether a trial was fair, disclosure was complete, custody was lawful, and harmed parties received remedy. We therefore score *events* on the realized path (H_t, R_t), the remedies actually enacted (M_t, J_t^v), and whether prior hazardous routes are *structurally* and *provably* closed (HZ = 0,

$S^* = 1$). Counting training hours without changed outcomes over-credits intentions.

30.2 Operational semantics (instantiation)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \,(\text{rights/procedure breach}) \,\text{or}\, H_t \geq h_{\min}\big) \,\text{and}\, C_t = 1.$$

$C_t = 1$ when an accountable body (police, prosecution service, court administration, judiciary where administrative, prisons/probation) had adequate knowledge/freedom and crossed a fault threshold (intent, recklessness, gross negligence). If $C_t = 0$ (accident), record H_t for learning but do not set $R_t = 1$.

Harm H_t. Transparent scales: unlawful custody days, appeal/writ success implying harm, collapsed trials from procedural fault, victim/witness detriment (safety, livelihood), verified financial loss, privacy harms.

Mercy/Justice M_t, J_t^v. $M_t = 1$ when restoration is *delivered*: release and record correction, compensation/ex gratia, re-hearing, special measures/support for victims/witnesses, privacy repair. $J_t^v = 1$ when *rectification/accountability* occurs: protocol/tool change, sanctions/discipline, public notice/regulatory engagement.

Near-misses and drills. Near-miss (late but pre-trial disclosure fixed before prejudice) is not $R_t = 1$; convert to *eventized drills* ($D_t =$

1): disclosure file-build simulations, custody-decision exercises, digital-evidence red-teams. Drills support S^* but do not mint M, J^v.

30.3 Evaluator and constraints (justice view; broken to fit)

$$\mathcal{J}_{\text{jus}} = \sum_{t\geq 0}\Big(\alpha\,\Delta L_t + \gamma\,\Delta F_t + \mu\,M_t + \nu\,J_t^{v}\Big) - \sum_{t\geq 0}\Big(\beta\,H_t + \kappa\,R_t\Big), \tag{30.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau\leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^* \in \{0,1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t credits *delivered* justice-care (special measures provided, safeguarding actions completed, timely communications); ΔF_t credits *exercised* rights (timely trial, access to counsel, appeals upheld, equality-of-arms remedies).

30.4 Minimal-trigger doctrine (system policy)

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (file build, digital search, custody thresholds, scheduling, interpreter routes, privacy controls), fix, tests, owner+deadline, publication plan (privacy-preserving).

2. **Remedy:** Enact M_t (release/correction/compensation/support) and J_t^v (rectification/accountability: protocol/tool change, sanctions, public guidance).

3. **Confirm:** Stronger-incentive reenactment: cohort replay → adversarial/stress drill → monitored live audits. On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats add cost (κ and H_t) but no new kind-level credit (typal/concave); one-off + closure dominates (Ch. 11).

30.5 Incident classes & scope locks (examples)

- **DISC-FAIL** — disclosure failure (unused material, digital search omissions, late service).
- **CASE-DEL** — excessive delay beyond standards causing harm (witness attrition, custody impact).
- **REMAND-EX** — excessive/unlawful remand (thresholds misapplied, reviews missed).
- **WCONV** — wrongful conviction (fresh evidence, process fault, unsafe directions).
- **EVID-LOSS** — evidence loss/spoliation (storage/chain failures).

- **INT-FAIL** — interpreter/translation failure affecting fairness.

- **SAFEG-WIT** — witness/victim safeguarding failure (intimidation, contact breaches).

- **DATA-PRIV** — privacy/data breach (case files, exhibits, open justice misapplication).

- **DIGI-EVID** — digital-evidence ingestion/search tool failure (scope, filters, audit).

- **CP-CHARGE** — charging decision error (over/under-charging contrary to code).

Each class requires a written *scope lock* freezing inputs, thresholds, tools, and preconditions; new causal vectors open a new class c' rather than relabeling repeats.

30.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Unlawful custody days; quashed/unsafe convictions; collapsed trials from procedure faults; victim/witness detriment; privacy harms
B_t (breach)	Judicial/tribunal findings; inspectorate/audit reports; code/protocol violations; custody review breaches
C_t (culpability)	Safety-case/process gap vs. standard; foreseeability; file/audit logs; scheduling/chain-of-custody records
M_t (mercy)	Release/record correction; compensation/ex gratia; special measures/support; privacy remediation
J_t^v (justice)	Protocol/tool change shipped; sanctions/discipline; practice directions; public guidance/notices
ΔL_t	Delivered support: special measures, safeguarding actions, timely comms, case progression reviews
ΔF_t	Rights exercised: timely trial, access to counsel, appeals upheld, equality-of-arms remedies
S_c^*	Passed cohort replay + adversarial/stress drills + monitored live audits; artefacts linked (privacy-preserving)

30.7 Confirmation tests (design pattern for justice)

Step 1: Cohort replay. Re-run matched cases (offence type, complexity, volume of digital evidence, custody status) showing removal of the prior vector.
Step 2: Adversarial/stress drills. Disclosure red-teams (unused material traps), digital-evidence search challenges, custody threshold edge-cases, interpreter stress tests; pre-registered pass metrics.
Step 3: Monitored live. Short live window with independent observers (inspectorate/ombudsman), random file audits, custody review sampling. On pass, set $S^*(c) = 1$.
Scope lock. Freeze class scope; later failures with different causality open c'.

30.8 Tool-gating and interim safeguards after a trigger

Until $S^*(c) = 1$: impose disclosure checklists with supervisor sign-off; halt risky custody extensions pending senior review; require early case management hearings; enable independent file audits; strengthen witness protection orders; restrict risky digital-evidence tools to approved workflows; privacy redaction defaults. Remove gates only after closure.

30.9 Micro-vignettes (worked examples)

(A) Disclosure failure (DISC-FAIL). *Trigger:* Late service of unused material; trial collapses; judicial finding of fault $\Rightarrow R_t = 1$. *Remedy:* Record correction; apology; compensation for costs; protocol/tool change (digital search scopes, audit trails), training with supervision $(M_t, J_t^y = 1)$.
Confirm: Cohort replay on matched case types; red-team disclosure traps; monitored live audits; $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Excessive remand (REMAND-EX). *Trigger:* Reviews missed; custody threshold misapplied; quashed detention $\Rightarrow R_t = 1$. *Remedy:* Immediate release; record correction; compensation; custody-decision checklist with senior sign-off $(M_t, J_t^y = 1)$.
Confirm: Replay custody decisions; stress edge-cases; monitored live sampling; closure on pass.

(C) Interpreter failure (INT-FAIL). *Trigger:* Unqualified interpreter; fairness compromised; appeal succeeds $\Rightarrow R_t = 1$.
Remedy: Re-hearing; accredited vendor list; booking gates; remuneration/penalties $(M_t, J_t^y = 1)$.
Confirm: Stress drills with rare languages; monitored live audits; class closed on pass.

30.10 Dashboards and metrics (ledger view)

- **Harm/recurrence:** unlawful custody days; disclosure-fault collapses; unsafe convictions; recurrence by class; time-to-

remedy; time-to-closure T_c^*.

- **Closure:** # classes with $S^* = 1$; median T_c^*; age/backlog of open dockets.

- **Culture (love):** special measures delivered; safeguarding actions; timely comms; case progression reviews (ΔL_t proxies).

- **Freedom:** timely trials; appeals upheld; equality-of-arms remedies; lawful custody reviews (ΔF_t).

30.11 Anti-gaming and integrity

- **Paper compliance.** Require artefacts that show *events* changed (on-time disclosure, lawful custody decisions), not only training logs.

- **Scope creep.** Lock class definitions; do not relabel repeats as "new" unless the causal vector differs.

- **Under-reporting.** Enable protected disclosures; random independent file audits; publish privacy-preserving summaries.

- **Equity.** Confirm closure holds across protected characteristics; report subgroup justice metrics in confirmation packs.

30.12 One-page checklist (drop-in for justice agencies)

Justice Incident → Confirmation Checklist

Trigger captured? case IDs, timestamps, file/audit logs, custody records, interpreter/vendor details

Gate set? disclosure checklists; custody senior sign-off; early case management; independent audits; privacy defaults

Docket opened? root cause, fix, tests, owner & deadline, publish plan

Remedy enacted? release/record correction; compensation/support; accountability actions

Tests passed? cohort replay ✓ adversarial/stress drills ✓ monitored live audits ✓

Scope locked? class frozen; distinct vectors filed as c'

Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined

Gates removed? only post-closure; rationale logged

30.13 Limits and open problems

- **Attribution.** Multi-actor chains (police, prosecution, courts, prisons/probation, vendors) complicate C_t; adopt shared-fault taxonomies.

- **Complex digital evidence.** Scale/format drift requires evolv-

ing tools and audits; pair model/tool updates with event-ledger confirmation.

- **Rare but severe harms.** Wrongful convictions are rare but catastrophic; heavy reliance on adversarial drills and independent review is needed while preserving honesty locks.

30.14 What this chapter contributes

A full translation of the minimal-trigger doctrine to justice & courts: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and evidence standards. With typal/concave credits and $\kappa > 0$, repeating the same vector adds cost without adding kind-level value; one-off and confirmed closure dominates. “‘□0■

Chapter 31

Casebook: Policing: Use of Force, Stops, and Confirmation

Notation

Domain mapping (frontline policing: call handling, patrol, investigations, custody, neighbourhoods, public order).
An *event* is a policing outcome with adjudication potential (stop/search, arrest, use of force, pursuit, custody care, disclosure, data handling).
$R_t \in \{0, 1\}$: **rejection** (culpable rights/safety breach or harm $\geq h_{\min}$); $B_t \in \{0, 1\}$: breach (law/code/standard: PACE, use-of-force policy, pursuit policy, equality, privacy, safeguarding); $C_t \in \{0, 1\}$: culpability gate (intent/reckless/gross negligence

by officer/unit/force/contractor); $H_t \geq 0$: realized harm (injury/death, unlawful deprivation of liberty, property/financial harm, privacy harm).
$M_t, J_t^v \in \{0, 1\}$: enacted mercy/justice (restoration to the actually harmed; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: increments to care/protection (love) and exercised rights/freedoms (freedom).
Illustrative incident classes $c \in C$: **UOF-HARM** use-of-force causing unlawful/avoidable harm; **STOP-FAIL** unlawful/disproportionate stop/search; **CUST-CARE** custody care failure (medical/suicide risk); **PURS-POL** pursuit policy breach; **DV-RISK** domestic abuse safeguarding failure; **DISC-POL** disclosure failure to CPS; **EVID-LOSS** evidence loss/spoliation; **DATA-PRIV** privacy/data breach (ANPR/body-worn/audit); **BIAS-DISP** disproportionality breach (pattern-level).
Switches: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0, 1\}$ per class. Evaluator $\mathcal{J}$ is event-valued (Ch. 9).

31.1 Why an event-valued ledger fits policing

Public trust hinges on *what actually happens*: were powers used lawfully and proportionately, were people kept safe in custody and during pursuits, were rights respected, and are repeat causal routes *structurally* and *provably* closed (HZ = 0, $S^* = 1$)? Counting training hours over-credits intentions. The ledger scores realized harms/breaches (H_t, R_t), enacted remedies (M_t, J_t^v), and proven

closure.

31.2 Operational semantics (instantiation)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \,(\text{rights/procedure breach}) \,\text{or}\, H_t \geq h_{\min}\big) \,\text{and}\, C_t = 1.$$

$C_t = 1$ where an accountable officer/unit/force/vendor had adequate knowledge/freedom and crossed a fault threshold (intent, recklessness, gross negligence) against a published *safety/ethics case*. If $C_t = 0$ (accident), record H_t for learning but do not set $R_t = 1$.

Harm H_t. Injury/fatality, unlawful detention minutes/days, unlawful property damage, documented trauma, privacy harms. Publish $h_{\min}$ (e.g., hospital-level injury, custody deprivation beyond threshold).

Mercy/Justice M_t, J_t^v. $M_t = 1$ when restoration is *delivered*: medical care, record correction/expungement, apology/compensation, property restoration. $J_t^v = 1$ when *rectification/accountability* is enacted: policy/process change, tool-gating, discipline/sanctions, referral to oversight, public notice.

Near-misses and drills. Averted collisions in pursuit, near-miss force applications caught by supervision, halted unlawful stop are not $R_t = 1$; convert to *eventized drills* ($D_t = 1$): scenario replays

with body-worn video (BWV), simulation ranges, custody drills. Drills support S^* but do not mint M, J^v.

31.3 Evaluator and constraints (policing view; broken to fit)

$$\mathcal{J}_{\text{pol}} = \sum_{t\geq 0}\Big(\alpha\,\Delta L_t + \gamma\,\Delta F_t + \mu\,M_t + \nu\,J_t^v\Big) - \sum_{t\geq 0}\Big(\beta\,H_t + \kappa\,R_t\Big), \tag{31.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau\leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^* \in \{0,1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t credits *delivered* protections (BWV on and available, medical checks, safeguarding actions, safe tactics); ΔF_t credits *exercised* rights (lawful grounds given, appeal upheld, records corrected, privacy rights fulfilled).

31.4 Minimal-trigger doctrine (force/authority policy)

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (tactics, tool, policy, supervision, roster, vendor), fix, tests, owner+deadline, publication plan (privacy-preserving).

2. **Remedy:** Enact M_t (care, apology/compensation, record correction) and J_t^v (rectification/accountability: tool-gates, policy/tactics change, discipline, referral).

3. **Confirm:** Stronger-incentive reenactment: cohort replay with BWV → stress simulations (night, crowd, weapon cues) → monitored live audits. On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats add cost (κ and H_t) but no new kind-level credit (typal/concave). One-off + closure dominates (Ch. 11).

31.5 Incident classes & scope locks (examples)

- **UOF-HARM** — avoidable injury/fatality from force (CEW/baton/restraint/dog/firearm) contrary to necessity/proportionality.
- **STOP-FAIL** — stop/search without lawful grounds or disproportional targeting beyond thresholds.
- **CUST-CARE** — custody risk assessment/observation/medical failure (self-harm, withdrawal, ligature).
- **PURS-POL** — pursuit breach (risk assessment, termination rules, PIT/TPAC misuse).
- **DV-RISK** — domestic abuse safeguarding failure (risk flags, protection orders, victim contact).

- **DISC-POL** — disclosure failure to prosecutors; late/omitted unused material.

- **EVID-LOSS** — evidence loss/spoliation (property handling, digital chain).

- **DATA-PRIV** — privacy/data breach (ANPR/BWV retention/access, unlawful processing).

- **BIAS-DISP** — statistically significant disproportionality not justified by exposure, with culpable governance gap.

Each class includes a written *scope lock* (tools, thresholds, environments, policies); new causal vectors open class c' rather than relabeling repeats.

31.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Medical/hospital records; injury/fatality reports; unlawful detention minutes/days; documented property/privacy harms
B_t (breach)	PACE/code/policy violations; pursuit logs; BWV/ANPR audits; inspectorate/ombudsman findings; disclosure audit fails
C_t (culpability)	Safety-case/process gap vs. standard; foreseeability; supervision/roster records; tool logs; training currency
M_t (mercy)	Medical care delivered; record correction/expungement; apology/compensation; property restoration
J_t^v (justice)	Policy/tactics change; tool-gates; discipline/sanctions; oversight referrals; vendor sanctions
ΔL_t	BWV usage compliance; safeguarding actions; de-escalation drills; custody observation/medical checks delivered
ΔF_t	Rights exercised: grounds stated; appeal upheld; data/access/erasure fulfilled; stop receipts issued
S_c^*	Passed BWV cohort replay + stress simulations + monitored live audits; artefacts linked (privacy-preserving)

31.7 Confirmation tests (design pattern for forces)

Step 1: BWV cohort replay. Re-sample matched incidents (time of day, call type, demographics, location risk) and show the prior vector is removed (e.g., force level, pursuit termination).
Step 2: Stress simulations. Night/crowd/weapon-cue sims; custody drills incl. medical withdrawal; red-team stops for lawful grounds; confirm policy/tool changes hold.
Step 3: Monitored live. Short live audits with independent observers (civilian panels/commissioners), random BWV review, ANPR/data audits; pre-registered pass metrics.
Scope lock. Freeze class scope; distinct causal vectors open c'.

31.8 Tool-gating and interim safeguards after a trigger

Until $S^*(c) = 1$: require BWV pre-activation for relevant encounters; supervisor approval for high-risk tools (CEW, TPAC manoeuvres); tightened pursuit termination thresholds; enhanced custody medical checks; stop/search oversight panels; privacy defaults on data tools. Remove gates only after closure.

31.9 Micro-vignettes (worked examples)

(A) Unlawful stop/search (STOP-FAIL). *Trigger:* Grounds not established; BWV shows inadequate suspicion; disproportional

pattern present $\Rightarrow R_t = 1$.
Remedy: Record correction/apology; compensation where due; policy refresh; supervisor gate on hotspots; panel oversight ($M_t, J_t^v = 1$).
Confirm: BWV cohort replay; red-team lawful/ulawful grounds tests; monitored live with disproportionality metrics; $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Custody care failure (CUST-CARE). *Trigger:* Missed risk assessment/observations; self-harm event; culpability confirmed $\Rightarrow R_t = 1$.
Remedy: Family support; record correction; medical protocol change; staffing/rota fix; cell environment changes ($M_t, J_t^v = 1$).
Confirm: Drills (withdrawal/ligature scenarios); monitored live custody audits; closure on pass.

(C) Pursuit collision (PURS-POL). *Trigger:* Policy breach on risk assessment/termination; collision with injury $\Rightarrow R_t = 1$.
Remedy: Compensation; pursuit policy/tool gates; driver re-certification; supervisor rules; vendor updates ($M_t, J_t^v = 1$).
Confirm: Simulator stress; controlled TPAC drills; monitored live with random audits; class closed on pass.

31.10 Dashboards and metrics (ledger view)

- **Harm/recurrence:** injury/fatality counts; unlawful detention time; recurrence by class; time-to-remedy; time-to-closure T_c^*.

- **Closure:** # classes with $S^* = 1$; median T_c^*; age/backlog of open dockets.

- **Culture (love):** BWV compliance; safeguarding actions; de-escalation drills; custody checks (ΔL_t proxies).

- **Freedom:** rights exercised (grounds/receipts); appeals upheld; privacy rights fulfilled; disproportionality trend improving (ΔF_t).

31.11 Anti-gaming and integrity

- **BWV gaps.** Enforce activation rules; random audits; sanctions for non-use; device health checks.

- **Metric laundering.** Publish distributions (force levels, stop hit-rates) not just means; pre-register pass criteria.

- **Scope creep.** Lock class definitions; do not relabel repeats as "new" unless causal vector differs.

- **Equity.** Confirm closure holds across protected characteristics; report subgroup disparities and improvements.

31.12 One-page checklist (drop-in for forces)

Policing Incident → Confirmation Checklist

Trigger captured? BWV, call/pursuit logs, custody/medical records, stop receipts
Gate set? BWV pre-activation; supervisor approval for high-risk tools; pursuit termination thresholds; custody medical checks
Docket opened? root cause, fix, tests, owner & deadline, publish plan
Remedy enacted? care/compensation; record correction; policy/tool change; accountability
Tests passed? BWV cohort replay ✓ stress simulations ✓ monitored live audits ✓
Scope locked? class frozen; distinct vectors filed as c'
Closure set? $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$; monitoring window defined
Gates removed? only post-closure; rationale logged

31.13 Limits and open problems

- **Attribution.** Multi-actor chains (officer, supervisor, control room, vendor) complicate C_t; adopt shared-fault taxonomies.
- **Rare but severe harms.** Death/serious injury require heavy reliance on simulations while preserving honesty locks (no

M, J^v for drills).

- **Adversarial drift.** Offender tactics and urban conditions evolve; require periodic re-confirmations for high-risk classes.

31.14 What this chapter contributes

A full translation of the minimal-trigger doctrine to policing: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and evidence standards. With typal/concave credits and $\kappa > 0$, repeating the same vector adds cost without adding kind-level value; one-off and confirmed closure dominates.

Chapter 32

Casebook: Prisons & Probation: Safety, Rehabilitation, and Confirmation

Notation

Domain mapping (custody and community): prisons, young offender institutions, immigration removal centres (custody-side); probation in the community, approved premises, parole/licence management (community-side).

An *event* is a custody/probation outcome with adjudication potential (assault/self-harm, healthcare incident, segregation, recall, licence breach, parole decision, MAPPA review, accom-

modation on release, data handling).

$R_t \in \{0,1\}$: **rejection** (culpable rights/safety breach or harm $\geq h_{\min}$); $B_t \in \{0,1\}$: breach (law/code/standard: use of force, segregation rules, healthcare duty, parole timeliness, recall lawfulness, equality, privacy); $C_t \in \{0,1\}$: culpability gate (intent/reckless/gross negligence by service/contractor/vendor); $H_t \geq 0$: realized harm (injury, death, unlawful deprivation of liberty, healthcare detriment, homelessness on release, privacy harm).

$M_t, J_t^v \in \{0,1\}$: enacted mercy/justice (restoration to the actually harmed; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: increments to care/protection (love) and exercised rights/freedoms (freedom).

Illustrative incident classes $c \in C$: **CUST-SAFE** assault/self-harm/suicide prevention failure; **MED-CARE** clinical care failure (medication, triage, emergency response); **SEG-RULE** segregation/isolation rules breach; **RECALL-ERR** unlawful or disproportionate recall; **PAROLE-DEL** parole delay/defect causing unlawful custody; **ACCOM-FAIL** no suitable accommodation at release; **LICENCE-FAIL** unsafe or unlawful licence conditions; **AP-INC** approved-premises safeguarding failure; **MAPPA-FAIL** public-protection plan failure; **DATA-PRIV** privacy/data breach.

Switches: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0,1\}$ per class. Evaluator $\mathcal{J}$ is event-valued (Ch. 9).

32.1 Why an event-valued ledger fits prisons & probation

Safety and rehabilitation hinge on *what actually happens*: were people kept safe, were healthcare and legal time-limits met, were release plans real, did licence conditions protect without overreach, and are repeat causal routes *structurally* and *provably* closed (HZ = 0, $S^* = 1$)? Counting policies or training over-credits intentions.

32.2 Operational semantics (instantiation)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \text{ (rights/standard breach) or } H_t \geq h_{\min}\big) \text{ and } C_t = 1.$$

$C_t = 1$ when an accountable prison/probation/contractor/vendor had adequate knowledge/freedom and crossed a fault threshold (intent, recklessness, gross negligence) relative to a published *safety/ethics case*. If $C_t = 0$ (accident), record H_t but do not set $R_t = 1$.

Harm H_t. Injury/fatality, self-harm with healthcare impact, unlawful custody days (parole delay/recall error), homelessness nights on release, loss of prescribed medication continuity, privacy harms. Publish $h_{\min}$.

Mercy/Justice M_t, J_t^v. $M_t = 1$ when *delivered* restoration occurs (medical care, record correction/release, accommodation provision, compensation/ex gratia, apology). $J_t^v = 1$ when *rectifica-*

tion/accountability is enacted (policy/process/tool changes; staffing floors; contractor sanctions; regulator/ombudsman engagements).

Near-misses and drills. Near-miss (ligature point noticed and removed pre-incident; recall halted pre-breach) is not $R_t = 1$; convert into *eventized drills* ($D_t = 1$): first-night safety drills, healthcare-code simulations, recall decision exercises. Drills support S^* but do not mint M, J^v.

32.3 Evaluator and constraints (custody/community view; broken to fit)

$$\mathcal{J}_{\text{pp}} = \sum_{t \geq 0} \Big(\alpha\, \Delta L_t + \gamma\, \Delta F_t + \mu\, M_t + \nu\, J_t^v \Big) - \sum_{t \geq 0} \Big(\beta\, H_t + \kappa\, R_t \Big), \tag{32.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau \leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^* \in \{0, 1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t credits *delivered* protections/support (ACCT/CSIP reviews completed, medication continuity, ligature removal, safe staffing, accommodation secured); ΔF_t credits *exercised* rights (lawful custody, timely parole, access to healthcare/legal, privacy rights, fair licence terms).

32.4 Minimal-trigger doctrine (service policy)

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (risk assessment, observations, clinical triage, segregation rules, recall decisioning, parole prep, accommodation pathways, vendor/tool), fix, tests, owner+deadline, publication plan (privacy-preserving).

2. **Remedy:** Enact M_t (care/release/record correction/accommodation/compensation) and J_t^v (rectification/accountability: policy/tool change, staffing floor, contractor sanctions, regulator notice).

3. **Confirm:** Stronger-incentive reenactment: cohort replay $\rightarrow$ stress drills (night/weekend/overcrowding spikes, complex health needs) $\rightarrow$ monitored live audits. On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats add cost (κ and H_t) but no new kind-level credit (typal/concave); one-off + closure dominates (Ch. 11).

32.5 Incident classes & scope locks (examples)

- **CUST-SAFE** — self-harm/suicide prevention failure; serious assault; force contrary to policy.

- **MED-CARE** — medication/triage failure; emergency response delay; continuity gaps across transfer/release.

- **SEG-RULE** — segregation beyond limits; unlawful isolation or reviews missed.

- **RECALL-ERR** — unlawful/disproportionate recall; evidence/threshold misapplied.

- **PAROLE-DEL** — parole dossier/board delay leading to unlawful custody days.

- **ACCOM-FAIL** — no suitable accommodation at release when duty/pathway existed.

- **LICENCE-FAIL** — licence conditions unsafe/unlawful (disproportionate, non-rehabilitative).

- **AP-INC** — approved premises safeguarding failure (violence/exploitation).

- **MAPPA-FAIL** — public-protection plan failure (handover gaps; information-sharing).

- **DATA-PRIV** — privacy/data breach (case files, health information, monitoring).

Each class must include a written *scope lock* (inputs, thresholds, tools, preconditions); new causal vectors open c'.

32.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Injury/fatality logs; healthcare records; unlawful custody days; homelessness nights at release; verified privacy harms
B_t (breach)	Policy/code violations; inspectorate/ombudsman findings; tribunal/judicial outcomes; clinical audit fails; segregation reviews missed
C_t (culpability)	Safety-case/process gap vs. standard; foreseeability; staffing/observation logs; clinical rosters; decision minutes
M_t (mercy)	Care delivered; record correction/release; accommodation provided; compensation/ex gratia; apology
J_t^v (justice)	Policy/process/tool change shipped; staffing floors; sanctions/contract remedies; regulator/ombudsman notifications
ΔL_t	Delivered protections: ACCT/CSIP reviews; medication continuity; ligature removal; safe staffing; approved-premises safeguards
ΔF_t	Rights exercised: timely parole; lawful recall; access to healthcare/legal; fair licence terms; data rights fulfilled
S_c^*	Passed cohort replay + stress drills +

32.7 Confirmation tests (design pattern for custody/probation)

Step 1: Cohort replay. Re-run matched cohorts (first-night custody, mental-health flags, recall types, parole-ready cases) and show removal of the prior vector.
Step 2: Stress drills. Night/weekend staffing; high-occupancy; complex comorbidities; transfers; recall decision edge-cases; accommodation scarcity drills; pre-registered pass metrics.
Step 3: Monitored live. Short live window with independent observers (inspectorate/IMB/probation inspectors), random file/cell audits, healthcare continuity checks. On pass, set $S^*(c) = 1$.
Scope lock. Freeze class scope; later failures with different causality open c'.

32.8 Tool-gating and interim safeguards after a trigger

Until $S^*(c) = 1$: raise observation levels; remove ligature points; clinical on-call floor; segregation time caps with senior sign-off; recall moratorium on borderline cases pending governance; accelerated parole file-build; default accommodation escalation for near-release; privacy defaults. Remove gates only after closure.

32.9 Micro-vignettes (worked examples)

(A) First-night self-harm (CUST-SAFE). *Trigger:* First-night risk assessment missed; self-harm with hospitalisation; culpability confirmed $\Rightarrow R_t = 1$.
Remedy: Medical care; family contact support; apology/compensation; staffing floor and assessment redesign; ligature audit ($M_t, J_t^v = 1$).
Confirm: Cohort replay of first-night cases; stress drills (night/weekend); monitored live audits; $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Unlawful recall (RECALL-ERR). *Trigger:* Evidence threshold misapplied; tribunal quashes recall; custody days accrued $\Rightarrow R_t = 1$.
Remedy: Release; record correction; compensation; decision-tool gates; training with senior review ($M_t, J_t^v = 1$).
Confirm: Replay recall types; red-team borderline cases; monitored live sampling; closure on pass.

(C) Parole delay (PAROLE-DEL). *Trigger:* Dossier delays cause unlawful custody days; governance fault found $\Rightarrow R_t = 1$.
Remedy: Expedited hearing; release/record correction; compensation; process/tooling fix; staffing reallocation ($M_t, J_t^v = 1$).
Confirm: Matched-case replay; stress queue spikes; monitored live; class closed on pass.

32.10 Dashboards and metrics (ledger view)

- **Harm/recurrence:** injury/fatality counts; unlawful custody days; homelessness nights at release; recurrence by class; time-to-remedy; T_c^*.
- **Closure:** # classes with $S^* = 1$; median T_c^*; backlog/age of open dockets.
- **Culture (love):** ACCT/CSIP completion; medication continuity; safe staffing; approved-premises safeguards (ΔL_t proxies).
- **Freedom:** timely parole; lawful recalls; fair licence conditions; privacy/data rights fulfilled (ΔF_t).

32.11 Anti-gaming and integrity

- **Under-reporting.** Protected disclosures; random file/cell audits; healthcare reconciliation; privacy-preserving public summaries.
- **Scope creep.** Lock class definitions; do not relabel repeats as "new" unless causal vector differs.
- **Equity.** Confirm closure across protected groups; publish subgroup metrics for harms and remedies.
- **Vendor accountability.** Healthcare/monitoring vendors must supply artefacts in confirmation packs; no black-box exemptions.

32.12 One-page checklist (drop-in for custody & probation)

Custody/Probation Incident → Confirmation Checklist

Trigger captured? case IDs, timestamps, risk/observation logs, clinical records, recall/parole minutes

Gate set? observation level raised; ligature removal; clinical on-call; recall moratorium; accelerated parole file-build; accommodation escalation

Docket opened? root cause, fix, tests, owner & deadline, publish plan

Remedy enacted? care/release; record correction; accommodation; compensation; accountability

Tests passed? cohort replay ✓ stress drills ✓ monitored live audits ✓

Scope locked? class frozen; distinct vectors filed as c'

Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined

Gates removed? only post-closure; rationale logged

32.13 Limits and open problems

- **Attribution.** Multi-actor chains (prison health provider, custody staff, probation, parole board, housing authority, vendors) complicate C_t; adopt shared-fault taxonomies.

- **Resource constraints.** Staffing/accommodation scarcity confound fixes; use T_c^* targets, triage by harm, and transparency.
- **Rare but severe harms.** Deaths in custody are low-frequency but catastrophic; rely on drills and independent investigation while preserving honesty locks (no M, J^v for drills).

32.14 What this chapter contributes

A full translation of the minimal-trigger doctrine to prisons & probation: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and evidence standards. With typal/concave credits and $\kappa > 0$, repeating the same vector adds cost without adding kind-level value; one-off and confirmed closure dominates.

Chapter 33

Casebook: NHS (Acute & Primary Care): Sentinel Events, Medication Safety, and Confirmation

Notation

Domain mapping (acute & primary care): emergency departments, wards/theatres/ICU, maternity/neonatal, community/GP practices, outpatients, ambulance interface, digital (EHR/e-prescribing).

An *event* is a patient-relevant outcome with adjudication potential (diagnosis/treatment/medication/surgery/handover/discharge/safeguarding/privacy).

$R_t \in \{0,1\}$: **rejection** (culpable rights/safety breach or harm $\geq h_{\min}$); $B_t \in \{0,1\}$: breach (law/code/standard: duty of candour, never-event policy, safeguarding, consent/records, infection control, medicines policy); $C_t \in \{0,1\}$: culpability gate (intent/reckless/gross negligence by trust/PCN/practice/vendor/contractor); $H_t \geq 0$: realized harm (death, severe/moderate harm, readmission, avoidable deterioration, privacy harm, unsafe discharge harm).

$M_t, J_t^v \in \{0,1\}$: enacted mercy/justice (restoration to the *actually* harmed; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: increments to care/protection (love) and exercised rights/freedoms (freedom).

Illustrative incident classes $c \in C$: **SURG-NEVER** wrong-site/procedure/patient surgery; **MED-ERR** medication error (dose/drug/route/interactions); **DIAG-DEL** diagnostic delay/miss (e.g. cancer/sepsis/MI/stroke); **SEPSIS-FAIL** failure to recognise/escalate; **DETERIOR-FAIL** observation/escalation failure (NEWS2); **DISCH-SAFE** unsafe discharge/transfer; **SAFE-CHILD** safeguarding failure (child/vulnerable adult); **IPC-BREAK** infection prevention/control breach; **MH-CRISIS** mental-health crisis pathway failure; **EHR-PRIV** privacy/data breach (records/e-prescribing).

Switches: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0,1\}$ per class. Evaluator $\mathcal{J}$ is event-valued (Ch. 9).

33.1 Why an event-valued ledger fits healthcare

Patients and families experience *events*: timely diagnoses, correct medicines, safe surgery, safe discharge, privacy kept. Counting policies, training hours, or nominal staffing over-credits intentions. We score realized outcomes (H_t, R_t), enacted remedies (M_t, J_t^v), and whether hazardous routes are *structurally* and *provably* closed (HZ = 0, $S^* = 1$).

33.2 Operational semantics (instantiation)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \,(\text{rights/standard breach}) \,\text{or}\, H_t \geq h_{\min}\big) \,\text{and}\, C_t = 1.$$

$C_t = 1$ when an accountable provider/vendor had adequate knowledge/freedom and crossed a fault threshold (intent, recklessness, gross negligence) relative to a published *safety case*. If $C_t = 0$ (accident), record H_t for learning but do not set $R_t = 1$.

Harm H_t. Transparent scales: death, severe/moderate harm (trust severity scale), unplanned ICU transfer, 7/30-day readmission, medicine-related ED visit, safeguarding harm, privacy harms. Publish $h_{\min}$.

Mercy/Justice M_t, J_t^v. $M_t = 1$ when restoration is *delivered*: clinical follow-up/repair, apology and duty-of-candour meeting,

compensation/ex gratia, medication/device replacement, safe re-discharge with supports. $J_t^v = 1$ when *rectification/accountability* is enacted: protocol/tool change, checklist, staffing floor, vendor patch, sanctions/regulator notice.

Near-misses and drills. Near-miss (barcode scan stops wrong drug; theatre check averts wrong site) is not $R_t = 1$; convert to *eventized drills* ($D_t = 1$): sepsis drills, time-out simulations, discharge role-plays, EHR red-team. Drills support S^* but do not mint M, J^v.

33.3 Evaluator and constraints (health view; broken to fit)

$$\mathcal{J}_{\text{nhs}} = \sum_{t\geq 0}\Big(\alpha\,\Delta L_t + \gamma\,\Delta F_t + \mu\,M_t + \nu\,J_t^v\Big) - \sum_{t\geq 0}\Big(\beta\,H_t + \kappa\,R_t\Big), \tag{33.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau\leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^* \in \{0,1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t credits *delivered* protections (time-outs, barcode scanning, rapid sepsis bundles, safe discharge plans). ΔF_t credits *exercised* rights (informed consent, second opinions, access/correction of records).

33.4 Minimal-trigger doctrine (trust/PCN policy)

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (teamwork/checklists, e-prescribing rules, lab/radiology turnaround, escalation ladders, discharge pathways, vendor tools), fix, tests, owner+deadline, publication plan (privacy-preserving).

2. **Remedy:** Enact M_t (clinical repair, apology/DoC, compensation, supports) and J_t^v (rectification/accountability: process/tool change, staffing floor, vendor patch, regulator notice).

3. **Confirm:** Stronger-incentive reenactment: cohort replay $\rightarrow$ stress drills (out-of-hours/handovers/peaks) $\rightarrow$ monitored live audits. On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats add cost (κ and H_t) but no new kind-level credit (typal/concave); one-off + closure dominates (Ch. 11).

33.5 Incident classes & scope locks (examples)

- **SURG-NEVER** — wrong-site/procedure/patient surgery; retained foreign object; wrong implant.

- **MED-ERR** — wrong drug/dose/route/omission; look-alike sound-alike (LASA); anticoagulant/insulin/opiate high-risk sets.
- **DIAG-DEL** — delayed/missed diagnosis (cancer, MI, stroke, PE, aortic catastrophe, AAA, cauda equina).
- **SEPSIS-FAIL** — failure to recognise/escalate (NEWS2, lactate, antibiotics/fluids delay).
- **DETERIOR-FAIL** — observation/early warning escalation failure; out-of-hours deterioration not acted upon.
- **DISCH-SAFE** — unsafe discharge: meds/reconciliation missing, equipment/care packages absent, transport risk, no safety-netting.
- **SAFE-CHILD** — safeguarding failure (missed non-accidental injury signals; MARAC/MASH gaps).
- **IPC-BREAK** — infection prevention/control breach (asepsis/isolation/PPE/environmental cleaning).
- **MH-CRISIS** — mental-health crisis pathway failure (risk assessment, ligature environment, 136 suite, community follow-up).
- **EHR-PRIV** — privacy/data breach (misdirected letters, open-access screens, unredacted sharing, e-prescribing exposure).

Each class requires a written *scope lock* (assets, thresholds, tools, handover points); new causal vectors open c'.

33.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Death/severe/moderate harm; unplanned ICU transfer; readmission within 7/30 days; medication-related ED visit; privacy harms
B_t (breach)	Never-event criteria met; duty-of-candour triggers; medicines policy/EHR audit fails; safeguarding code breaches; IPC audit fails
C_t (culpability)	Safety-case gap vs. standard; foreseeability; checklist/observation logs; e-prescribing overrides; handover records
M_t (mercy)	Clinical repair/follow-up; apology/DoC meeting; compensation/ex gratia; medication/devices replaced; safe re-discharge supports
J_t^v (justice)	Protocol/tool change shipped; checklist added; staffing floor; vendor patch; regulator/ombudsman notices
ΔL_t	Protections delivered: time-outs, barcode scanning, sepsis bundles, observation/NEWS2 escalations, discharge safety-netting
ΔF_t	Rights exercised: informed consent, second opinions, information/access/correction of records
S_c^*	Passed cohort replay + stress drills + monitored live audits; artefacts linked

33.7 Confirmation tests (design pattern for trusts/PCNs)

Step 1: Cohort replay. Re-run matched cases (same lists/clinics/wards, acuity, out-of-hours mix) showing removal of the prior vector; use SPC or matched controls.
Step 2: Stress drills. Out-of-hours, handover, peak ED arrivals, staff sickness, mixed polypharmacy; simulate barcode failures/EHR downtime; pre-register pass metrics.
Step 3: Monitored live. Short live window with independent observers (clinical governance/ICB/inspectorate), random chart/pharmacist audits, sepsis/NEWS2 spot checks. On pass, set $S^*(c) = 1$.
Scope lock. Freeze class scope; different causal vectors open c'.

33.8 Tool-gating and interim safeguards after a trigger

Until $S^*(c) = 1$: second-check pharmacist for high-risk meds; e-prescribing hard-stops and barcode scanning mandatory; surgical time-out with senior sign-off; NEWS2 escalation floors; sepsis alerts; discharge checklists with safety-netting; safeguarding senior review; privacy redaction defaults; vendor change-control freezes. Remove gates only after closure.

33.9 Micro-vignettes (worked examples)

(A) Wrong-site surgery (SURG-NEVER). *Trigger:* Time-out deviation and marking error lead to wrong site; never-event criteria met $\Rightarrow R_t = 1$.
Remedy: Clinical repair where possible; duty-of-candour meeting; compensation; time-out redesign (two-voice check; imaging on screen); instrument/marking policy; vendor tray labelling ($M_t, J_t^v = 1$).
Confirm: Theatre list replay; out-of-hours drills; monitored live observers; $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Sepsis recognition delay (SEPSIS-FAIL). *Trigger:* NEWS2/lactate not acted on; antibiotics delayed; ICU transfer; culpability confirmed $\Rightarrow R_t = 1$.
Remedy: Family meeting; clinical follow-up; sepsis bundle timer/alerts; escalation ladder; ward pharmacist review ($M_t, J_t^v = 1$).
Confirm: Cohort replay; ED/ward peak stress; monitored live bundle-time audits; closure on pass.

(C) Unsafe discharge (DISCH-SAFE). *Trigger:* Meds reconciliation missing; equipment/care package absent; readmission with harm $\Rightarrow R_t = 1$.
Remedy: Safe re-discharge with supports; transport fix; equipment delivered; GP/community handover; checklist/tooling updates ($M_t, J_t^v = 1$).
Confirm: Matched discharge cohorts; weekend stress; monitored live safety-netting calls; class closed on pass.

33.10 Dashboards and metrics (ledger view)

- **Harm/recurrence:** never-events; severe/moderate harm; unplanned ICU transfers; 7/30-day readmissions; med-error ED visits; recurrence by class; time-to-remedy; T_c^*.
- **Closure:** # classes with $S^* = 1$; median T_c^*; backlog/age of open dockets.
- **Culture (love):** time-out/barcode compliance; sepsis bundle times; discharge safety-netting delivered; safeguarding actions (ΔL_t).
- **Freedom:** consent completion; second-opinion access; information/records rights fulfilled (ΔF_t).

33.11 Anti-gaming and integrity

- **Paper compliance.** Require evidence that *events* changed (bundle times, error rates, readmissions), not only policy updates.
- **Scope creep.** Lock class definitions; do not relabel repeats as "new" unless causal vector differs.
- **Under-reporting.** Protected disclosures; random pharmacist/chart audits; privacy-preserving public summaries.
- **Vendor accountability.** EHR/e-prescribing vendors must supply artefacts (rules, logs) in confirmation packs; no black-box exemptions.

33.12 One-page checklist (drop-in for acute & primary care)

Healthcare Incident → Confirmation Checklist

Trigger captured? patient IDs, times, observations/labs, e-prescribing logs, handover notes

Gate set? second-check pharmacist; barcode/e-prescribing hard-stops; sepsis alerts; escalation floors; discharge checklist

Docket opened? root cause, fix, tests, owner & deadline, publish plan

Remedy enacted? clinical repair; duty-of-candour; supports/compensation; accountability actions

Tests passed? cohort replay ✓ stress drills ✓ monitored live audits ✓

Scope locked? class frozen; distinct vectors filed as c'

Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined

Gates removed? only post-closure; rationale logged

33.13 Limits and open problems

- **Attribution.** Multi-actor chains (ED/ward/theatre/pharmacy/community/ve complicate C_t; adopt shared-fault taxonomies.

- **Resource constraints.** Staffing/bed/kit scarcity confounds fixes; prioritise by harm and publish T_c^* trajectories.

- **Rare but severe harms.** Never-events are low-frequency but catastrophic; rely on drills and independent review while preserving honesty locks (no M, J^{ν} for drills).

33.14 What this chapter contributes

A full translation of the minimal-trigger doctrine to NHS acute & primary care: precise triggers (R_t), docketed remedies (M, J^{ν}), structural closure (S^*), and evidence standards. With typal/concave credits and $\kappa > 0$, repeating the same vector adds cost without adding kind-level value; one-off and confirmed closure dominates.

Chapter 34

Casebook: Mental Health: Crisis, Continuity, and Safeguarding

Notation

Domain mapping (mental health across settings): community/CMHT, crisis resolution/home treatment (CRHT), liaison psychiatry (ED/inpatient), crisis lines, CAMHS → AMHS transitions, perinatal, inpatient MH wards/136 suites, secure/forensic, and third-sector partners.

An *event* is a patient-/carer-relevant outcome with adjudication potential (crisis response, risk assessment, admission/discharge, medication safety, restraint/seclusion, safeguarding, aftercare, privacy).

$R_t \in \{0,1\}$: **rejection** (culpable rights/safety breach or harm $\geq h_{\min}$); $B_t \in \{0,1\}$: breach (law/code/standard: crisis-response SLAs, consent/least-restrictive principle, seclusion rules, meds-monitoring, safeguarding, aftercare duties, privacy/data);
$C_t \in \{0,1\}$: culpability gate (intent/reckless/gross negligence by trust/ICB/provider/vendor); $H_t \geq 0$: realized harm (self-harm/suicide/attempt, physical injury, unlawful deprivation of liberty, avoidable deterioration, homelessness/placement harm, privacy harm).
$M_t, J_t^v \in \{0,1\}$: enacted mercy/justice (restoration to the actually harmed; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: increments to care/protection (love) and exercised rights/freedoms (freedom).

Illustrative incident classes $c \in C$: **CRISIS-DEL** crisis response delay/failure; **RISK-FAIL** risk assessment/escalation failure; **136-FAIL** s136 pathway/holding breach; **RESTR-HARM** restraint/seclusion harm contrary to policy; **MED-SAFE** medication safety/monitoring failure (e.g. clozapine/lithium); **DC-AFTER** unsafe discharge/aftercare (incl. s117) failure; **TRANS-GAP** CAMHS→AMHS transition gap; **BED-OAP** out-of-area placement harm; **SAFE-CHILD** safeguarding failure (child/vulnerable adult); **EHR-PRIV** privacy/data breach. Switches: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0,1\}$ per class. Evaluator $\mathcal{J}$ is event-valued (Ch. 9).

34.1 Why an event-valued ledger fits mental health

Mental health assurance must answer *what actually happened*: did the crisis team arrive within agreed windows; was a safe, least-restrictive option provided; were seclusion rules followed; were lithium levels monitored; did the patient receive lawful aftercare; were children safeguarded; did a breach recur? Counting training, posters, or plans over-credits *capacity*. Our ledger scores real events (H_t, R_t), enacted restoration/rectification (M_t, J_t^v), and whether prior hazardous routes are *structurally* and *provably* closed (HZ = 0, $S^* = 1$).

34.2 Operational semantics (instantiation)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \,(\text{rights/standard breach}) \,\text{or}\, H_t \geq h_{\min}\big) \,\text{and}\, C_t = 1.$$

$C_t = 1$ when an accountable provider/vendor had adequate knowledge/freedom and crossed a fault threshold (intent, recklessness, gross negligence) relative to a published *safety/ethics case*. If $C_t = 0$, log H_t for learning but do not set $R_t = 1$.

Harm H_t. Self-harm/suicide or attempt; injury from unlawful/avoidable restraint; unlawful deprivation of liberty; avoidable deterioration (missed crisis window, failed escalation); homelessness nights after unsafe discharge; privacy harms. Publish $h_{\min}$.

Mercy/Justice M_t, J_t^v**.** $M_t = 1$ when *delivered*: clinical repair/aftercare, apology and family meeting, compensation/ex gratia, safe placement/accommodation, record correction. $J_t^v = 1$ when *rectification/accountability* is enacted: process/tool change, staffing floors, vendor/system patch, sanctions/regulator notice.

Near-misses & drills. Near-miss (escalation caught just in time; barcode stops wrong med) is not $R_t = 1$; convert to *eventized drills* ($D_t = 1$): crisis call-flow simulations, s136 pathway drills, restraint & ligature simulations, lithium/clozapine monitoring drills, discharge/aftercare role-plays. Drills support S^* but do not mint M, J^v.

34.3 Evaluator and constraints (mental health view; broken to fit)

$$\mathcal{J}_{\text{mh}} = \sum_{t\geq 0} \Big(\alpha\,\Delta L_t + \gamma\,\Delta F_t + \mu\,M_t + \nu\,J_t^v\Big) - \sum_{t\geq 0} \Big(\beta\,H_t + \kappa\,R_t\Big), \tag{34.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau\leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^* \in \{0,1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t credits *delivered* protections/supports (timely crisis attendance, safety plans, family liaison, therapeutic observations, safe environments). ΔF_t credits *exercised* rights (least-restrictive option, informed consent/advance statements, lawful aftercare, privacy/data rights).

34.4 Minimal-trigger doctrine (trust/ICB policy)

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (call triage, roster/coverage, handover to liaison, bed-flow/alternatives, seclusion suite rules, meds-monitoring, discharge/aftercare pathways, safeguarding, vendor tools), fix, tests, owner+deadline, publication plan (privacy-preserving).

2. **Remedy:** Enact M_t (clinical repair/aftercare/accommodation/supports, apology/compensation) and J_t^v (rectification/accountability: process/tool change, staffing floor, vendor patch, sanctions/regulator notice).

3. **Confirm:** Stronger-incentive reenactment: cohort replay $\rightarrow$ stress drills (night/weekend spikes, ED surges, high acuity) $\rightarrow$ monitored live audits. On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats add cost (κ and H_t) but no new kind-level credit (typal/concave); one-off + closure dominates (Ch. 11).

34.5 Incident classes & scope locks (examples)

- **CRISIS-DEL** — crisis team response delayed beyond standard causing harm; call triage/escalation failed.
- **RISK-FAIL** — risk assessment/observation/escalation failure (community, ward, or 136) leading to harm.
- **136-FAIL** — s136 pathway breach: place of safety unavailable; unlawful holding; assessments delayed beyond rules.
- **RESTR-HARM** — unlawful/avoidable restraint or seclusion harms contrary to policy/least-restrictive principle.
- **MED-SAFE** — medication/monitoring failure: clozapine (ANC), lithium (levels/renal/thyroid), antipsychotic ECG, rapid tranquilisation protocols.
- **DC-AFTER** — unsafe discharge/aftercare (incl. s117) with readmission/self-harm/homelessness.
- **TRANS-GAP** — CAMHS→AMHS transition gap with deterioration or loss to follow-up.
- **BED-OAP** — out-of-area placement with avoidable harm (distance, family separation, delayed reviews).
- **SAFE-CHILD** — safeguarding failure (exploitation/domestic abuse risk not escalated).
- **EHR-PRIV** — privacy/data breach (notes, letters, portals, observation data).

Each class includes a written *scope lock* (assets, thresholds, tools, preconditions); new causal vectors open c'.

34.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Self-harm/suicide/attempt; injury from restraint; unlawful deprivation of liberty minutes/days; avoidable deterioration; homelessness/readmission; privacy harms
B_t (breach)	Crisis/assessment timeliness breaches; least-restrictive/seclusion rule breaches; monitoring protocol breaches; aftercare/transition duty failures; privacy/code breaches
C_t (culpability)	Safety-case gap vs. standard; foreseeability; roster/call logs; observation/observation-interval records; medication-monitoring logs; discharge/aftercare minutes
M_t (mercy)	Clinical repair; apology/family meeting; compensation/ex gratia; accommodation/supports; record correction
J_t^v (justice)	Process/tool change shipped; staffing floors; vendor patches; sanctions/discipline; regulator/ombudsman notices
ΔL_t	Delivered protections: timely crisis attendance; safety plans; therapeutic observations; family liaison; safe environments
ΔF_t	Rights exercised: least-restrictive

34.7 Confirmation tests (design pattern for MH providers)

Step 1: Cohort replay. Re-run matched cohorts (same acuity/age/diagnoses, similar call windows, ED liaison loads) showing removal of the prior vector; include matched controls or SPC.
Step 2: Stress drills. Night/weekend spikes; simultaneous calls; ED surges; rapid tranquilisation scenarios; seclusion-rule edge cases; lithium/clozapine monitoring drills; pre-register pass metrics.
Step 3: Monitored live. Short live window with independent observers (clinical governance/ICB/inspectorate), random chart/observation/meds audits, family liaison checks. On pass, set $S^*(c) = 1$.
Scope lock. Freeze class scope; later failures with different causality open c'.

34.8 Tool-gating and interim safeguards after a trigger

Until $S^*(c) = 1$: crisis response *floor* times; senior clinician callbacks; enhanced observations; restraint/seclusion supervisor sign-off; safe-room/ligature audits; lithium/clozapine *hard-stops*; default aftercare bookings; accommodation escalation pathways; privacy redaction defaults; vendor change-control freezes. Remove gates only after closure.

34.9 Micro-vignettes (worked examples)

(A) Crisis response delay (CRISIS-DEL). *Trigger:* Multiple calls; team attendance beyond window; patient attempts self-harm; culpability confirmed $\Rightarrow R_t = 1$.
Remedy: Clinical repair; apology/family meeting; roster redesign (surge rules); escalation ladder; crisis-line integration ($M_t, J_t^v = 1$).
Confirm: Cohort replay on peak windows; night/weekend drills; monitored live; $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Restraint harm (RESTR-HARM). *Trigger:* Seclusion/restraint contrary to policy; injury occurs; observation gaps $\Rightarrow R_t = 1$.
Remedy: Medical care; apology/compensation; de-escalation training with supervisor gates; environment changes; vendor/device checks ($M_t, J_t^v = 1$).
Confirm: Simulated high-acuity drills; monitored live with random BWV/observation audits where available; closure on pass.

(C) Lithium monitoring failure (MED-SAFE). *Trigger:* Levels not taken; toxicity; ED admission; governance faults $\Rightarrow R_t = 1$.
Remedy: Clinical repair; medication review; lab/EHR alerts; dispensing locks; patient-held monitoring cards ($M_t, J_t^v = 1$).
Confirm: Cohort replay; stress drills with clinic backlogs; monitored live audits; class closed on pass.

34.10 Dashboards and metrics (ledger view)

- **Harm/recurrence:** self-harm/suicide/attempts; restraint/seclusion harms; unlawful holding time; unsafe discharge/readmission; recurrence by class; time-to-remedy; T_c^*.
- **Closure:** # classes with $S^* = 1$; median T_c^*; backlog/age of open dockets.
- **Culture (love):** crisis attendance floors; observation compliance; safety-plan completion; family liaison (ΔL_t proxies).
- **Freedom:** least-restrictive care; consent/advance statements recorded; lawful aftercare; privacy/data rights fulfilled (ΔF_t).

34.11 Anti-gaming and integrity

- **Paper compliance.** Require *event* change (response times, observation adherence, monitoring completion, readmissions), not only training logs.
- **Scope creep.** Lock class definitions; do not relabel repeats as "new" unless causal vector differs.
- **Under-reporting.** Protected disclosures; random chart/observation/meds audits; privacy-preserving public summaries.
- **Equity.** Confirm closure holds across age/sex/ethnicity/diagnosis; report subgroup metrics in confirmation packs.

34.12 One-page checklist (drop-in for MH services)

Mental Health Incident → Confirmation Checklist

Trigger captured? call/attendance times, risk/observation logs, meds-monitoring, seclusion records, discharge/aftercare minutes

Gate set? crisis floor times; supervisor sign-off for restraint/seclusion; lithium/clozapine hard-stops; aftercare bookings; accommodation escalation

Docket opened? root cause, fix, tests, owner & deadline, publish plan

Remedy enacted? clinical repair; apology/supports/compensation; record correction; accountability actions

Tests passed? cohort replay ✓ stress drills ✓ monitored live audits ✓

Scope locked? class frozen; distinct vectors filed as c'

Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined

Gates removed? only post-closure; rationale logged

34.13 Limits and open problems

- **Attribution.** Multi-actor chains (crisis lines, police/ambulance, ED liaison, wards, community teams, social

care, housing, vendors) complicate C_t; adopt shared-fault taxonomies.

- **Resource constraints.** Bed/roster pressures confound fixes; use T_c^* targets, harm-based triage, and transparent surge rules.
- **Rare but severe harms.** Low-frequency catastrophic events necessitate heavy reliance on drills while preserving honesty locks (no M, J^v for drills).

34.14 What this chapter contributes

A full translation of the minimal-trigger doctrine to mental health: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and evidence standards. With typal/concave credits and $\kappa > 0$, repeating the same vector adds cost without adding kind-level value; *one-off and confirmed closure dominates*.

Chapter 35

Casebook: Adult Social Care - Visits, MAR, Confirmation

Notation

Domain mapping (adult social care): local-authority commissioning, domiciliary/home care, residential/nursing homes, supported living, day services, direct payments/Personal Assistants (PAs), equipment/adaptations, and multi-agency safeguarding. An *event* is a person-relevant outcome with adjudication potential (care visit delivered/missed/unsafe, safeguarding breach, medication administration, falls/pressure injuries, discharge/transfer, accommodation/equipment, privacy/data).

$R_t \in \{0, 1\}$: **rejection** (culpable rights/safety breach or

harm $\geq h_{\min}$); $B_t \in \{0,1\}$: breach (law/code/standard: safeguarding duties, visit/plan adherence, MAR/meds policy, MCA/LPS[1] assessments, data protection); $C_t \in \{0,1\}$: culpability gate (intent/reckless/gross negligence by provider/LA/commissioner/contractor); $H_t \geq 0$: realized harm (injury, dehydration/malnutrition, pressure injury, missed medicines, hospital admission, unlawful restriction, homelessness/unsafe housing, privacy harm).

$M_t, J_t^v \in \{0,1\}$: enacted mercy/justice (restoration to the *actually* harmed; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: increments to care/protection (love) and exercised rights/freedoms (freedom).

Illustrative incident classes $c \in C$: **VISIT-MISS** missed/curtailed essential care visit; **SAFEG-ABUSE** abuse/neglect safeguarding failure; **MEDS-ADM** medication administration error (MAR); **FALLS-PLAN** falls risk plan failure; **PRESS-ULC** pressure ulcer prevention failure; **HYDR-NUTR** hydration/nutrition failure; **MCA-LPS** capacity/consent/LPS process failure; **DISCH-TRAN** unsafe hospital $\rightarrow$ care transfer; **EQUIP-ADPT** equipment/adaptation failure; **DATA-PRIV** privacy/data breach.

Switches: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0,1\}$ per class. Evaluator $\mathcal{J}$ is event-valued (Ch. 9).

[1] Mental Capacity Act (MCA) and Liberty Protection Safeguards (LPS).

35.1 Why an event-valued ledger fits adult social care

People feel outcomes, not policy binders: was the visit on time and complete; were meds given; was mum hydrated; did a pressure sore recur; was consent respected; did the hazard route *actually* close? Counting rosters, trainings, or portal features over-credits *capacity*. We score realized events (H_t, R_t), enacted remedies (M_t, J_t^y), and structural/proven closure (HZ = 0, $S^* = 1$).

35.2 Operational semantics (instantiation)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \text{ (rights/standard breach) or } H_t \geq h_{\min}\big) \text{ and } C_t = 1.$$

$C_t = 1$ when an accountable provider/LA/contractor had adequate knowledge/freedom and crossed a fault threshold (intent, recklessness, gross negligence) against a published *safety/ethics case*. If $C_t = 0$ (accident), record H_t for learning but do not set $R_t = 1$.

Harm H_t. Transparent scales: falls with injury, pressure injury (grade 2–4), dehydration/malnutrition markers, missed-time of essential personal care, missed MAR doses for high-risk meds, avoidable hospital admission, unlawful restriction, privacy harms. Publish $h_{\min}$.

Mercy/Justice M_t, J_t^v. $M_t = 1$ when *delivered* restoration occurs: catch-up care hours; rehydration/nutrition plan delivered; clinical follow-up; replacement kit/adaptations; apology/compensation; record correction. $J_t^v = 1$ when rectification/accountability is *enacted*: roster redesign, digital visit verification (DVE), staffing floors, escalation paths, sanctions/vendor fixes, regulator notices.

Near-misses & drills. Near-miss (late but completed visit; barcode/telecare alert averts harm) is not $R_t = 1$; convert to *eventized drills* ($D_t = 1$): unannounced call-ins, weekend roster drills, MAR simulations, transfer role-plays. Drills support S^* but do not mint M, J^v.

35.3 Evaluator and constraints (social care view; broken to fit)

$$\mathcal{J}_{\text{sc}} = \sum_{t\geq 0}\Big(\alpha\,\Delta L_t + \gamma\,\Delta F_t + \mu\,M_t + \nu\,J_t^v\Big) - \sum_{t\geq 0}\Big(\beta\,H_t + \kappa\,R_t\Big), \tag{35.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau\leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^* \in \{0,1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t credits *delivered* protections (visits delivered in full, hydration prompts, repositioning, equipment installed). ΔF_t credits *exercised* freedoms (consent/MCA respected, LPS lawfully applied, choice of routines/providers, data rights).

35.4 Minimal-trigger doctrine (commissioner/provider policy)

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (roster/route, DVE gaps, MAR controls, hydration/falls plans, equipment supply, MCA/LPS workflow, discharge coordination, data handling), fix, tests, owner+deadline, publication plan (privacy-preserving).

2. **Remedy:** Enact M_t (catch-up care, clinical follow-up, replacement kit, accommodation fixes, apology/compensation) and J_t^v (rectification/accountability: staffing/route redesign, DVE, sanctions/vendor patch, regulator notices).

3. **Confirm:** Stronger-incentive reenactment: cohort replay $\rightarrow$ weekend/peak stress drills $\rightarrow$ unannounced spot checks/monitored live. On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats add cost (κ and H_t) but no new kind-level credit (typal/concave). One-off + closure dominates (Ch. 11).

35.5 Incident classes & scope locks (examples)

- **VISIT-MISS** — missed/curtailed essential visit (washing, toileting, meals, meds prompts).
- **SAFEG-ABUSE** — abuse/neglect safeguarding failure (partnered with LA oversight).
- **MEDS-ADM** — MAR/med administration error (omission/wrong dose/time/high-risk meds).
- **FALLS-PLAN** — falls plan absent/ignored; repeat falls with injury.
- **PRESS-ULC** — failure of repositioning/skin checks → pressure injury.
- **HYDR-NUTR** — dehydration/malnutrition from care-plan failure.
- **MCA-LPS** — capacity not assessed; unlawful restriction; LPS paperwork/process failure.
- **DISCH-TRAN** — unsafe hospital discharge/transfer (no meds/equipment/care package).
- **EQUIP-ADPT** — kit/adaptation not provided/maintained (hoists, rails, ramps, commodes).
- **DATA-PRIV** — privacy/data breach (care notes/photos/portal access).

Each class requires a *scope lock* (inputs, thresholds, tools, preconditions); new causal vectors open c'.

35.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Injury; pressure ulcer grade 2–4; dehydration/malnutrition indicators; missed essential-care hours; hospital admission attributable to care failure; privacy harms
B_t (breach)	Safeguarding findings; MAR audit fails; visit/DVE logs; MCA/LPS breaches; discharge policy failures; equipment/adaptation standards not met; data protection audit fails
C_t (culpability)	Safety-case/process gap vs. standard; foreseeability; roster/route logs; MAR sheets; equipment orders/maintenance; MCA/LPS minutes; discharge/transfer notes
M_t (mercy)	Catch-up care hours; clinical follow-up; hydration/nutrition plans; replacement kit; accommodation fixes; apology/compensation
J_t^v (justice)	Roster redesign; DVE deployment; staffing floors; sanctions/vendor remedies; regulator/ombudsman notices
ΔL_t	Delivered protections: visits on time/complete; repositioning logs; hydration prompts; kit installed/maintained
ΔF_t	Rights exercised: consent recorded; MCA assessments; LPS lawfully ap-

35.7 Confirmation tests (design pattern for providers/LAs)

Step 1: Cohort replay. Re-run matched packages (same needs/visit frequency, geography, time-of-day) showing the prior vector removed; include matched controls.
Step 2: Stress drills. Weekend/evening spikes; winter illness/staff sickness; bank holidays; mixed MAR/kit needs; pre-register pass metrics.
Step 3: Unannounced/monitored live. Unannounced spot checks; GPS/DVE reconciliations; family confirmations; independent observers. On pass, set $S^*(c) = 1$.
Scope lock. Freeze class scope; later failures with different causality open c'.

35.8 Tool-gating and interim safeguards after a trigger

Until $S^*(c) = 1$: DVE mandatory; minimum visit lengths; no-lone-working where required; senior sign-off for roster compression; hydration/falls/skin *floors* (non-optional tasks); equipment loan fast-track; discharge moratorium without essentials; privacy redaction defaults; vendor change-control freezes. Remove gates only after closure.

35.9 Micro-vignettes (worked examples)

(A) Missed essential visit (VISIT-MISS). *Trigger:* Multiple missed morning visits; dehydration/UTI admission; DVE/roster faults $\Rightarrow R_t = 1$.
Remedy: Catch-up care; clinical follow-up; apology/compensation; route redesign; DVE deployment; staffing floor $(M_t, J_t^v = 1)$.
Confirm: Cohort replay; weekend stress; unannounced checks; $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Medication omission (MEDS-ADM). *Trigger:* MAR omissions for anticoagulant/insulin; harm occurs; governance gap confirmed $\Rightarrow R_t = 1$.
Remedy: Clinical repair; MAR/eMAR hard-stops; double-sign for high-risk meds; pharmacist input $(M_t, J_t^v = 1)$.
Confirm: MAR drills; monitored live spot audits; closure on pass.

(C) Unlawful restriction (MCA-LPS). *Trigger:* Restriction without capacity assessment/LPS; rights breach $\Rightarrow R_t = 1$.
Remedy: Record correction; lawful process; advocacy; staff training with supervisor gates $(M_t, J_t^v = 1)$.
Confirm: Case replay; stress drills with complex cases; unannounced audits; class closed on pass.

35.10 Dashboards and metrics (ledger view)

- **Harm/recurrence:** missed-visit hours; falls with injury; pressure injuries; MAR omissions; recurrence by class; time-

to-remedy; T_c^*.

- **Closure:** # classes with $S^* = 1$; median T_c^*; backlog/age of open dockets.

- **Culture (love):** visits on time/complete; hydration/repositioning delivered; kit installed/maintained (ΔL_t).

- **Freedom:** consent/MCA recorded; lawful LPS; choice respected; data rights fulfilled (ΔF_t).

35.11 Anti-gaming and integrity

- **Clock-in gaming.** Reconcile DVE GPS/telephony with family confirmations; random call-backs; sanctions for falsification.

- **Metric laundering.** Publish distributions (visit lengths, lateness) not just averages; pre-register pass criteria.

- **Scope creep.** Lock class definitions; do not relabel repeats as "new" unless causal vector differs.

- **Equity.** Confirm closure holds across disability/age/ethnicity; report subgroup metrics in confirmation packs.

- **Vendor accountability.** Require eMAR/DVE vendors to supply logs in confirmation packs; no black-box exemptions.

35.12 One-page checklist (drop-in for providers/commissioners)

Social Care Incident → Confirmation Checklist

Trigger captured? DVE logs, MAR sheets, roster/route, care-plan, equipment orders, MCA/LPS minutes

Gate set? DVE mandatory; minimum visit lengths; no-lone-working; hydration/falls/skin task floors; discharge essentials gate

Docket opened? root cause, fix, tests, owner & deadline, publish plan

Remedy enacted? catch-up care; clinical follow-up; kit/adaptations; apology/compensation; accountability

Tests passed? cohort replay ✓ weekend/peak drills ✓ unannounced spot checks/monitored live ✓

Scope locked? class frozen; distinct vectors filed as c'

Closure set? $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$; monitoring window defined

Gates removed? only post-closure; rationale logged

35.13 Limits and open problems

- **Workforce constraints.** Scarcity and churn confound fixes; use T_c^* targets, harm-based triage, and transparent escalation to alternate providers.

- **Multi-actor attribution.** Providers, LAs, NHS, landlords, vendors: adopt shared-fault taxonomies for C_t.
- **Hidden harms.** Social isolation and dignity losses require qualitative panels alongside quantitative KPIs.

35.14 What this chapter contributes

A full translation of the minimal-trigger doctrine to adult social care: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and evidence standards. With typal/concave credits and $\kappa > 0$, repeating the same vector adds cost without adding kind-level value; one-off and confirmed closure dominates. “‘□0■

Chapter 36

Casebook: Child Protection & Safeguarding: Multi-Agency Hazards and Confirmation

Notation

Domain mapping (children's services end-to-end): MASH/Front Door triage, Early Help, s47 enquiries, strategy meetings, Child Protection (CP) conferences & plans, Looked-After Children (LAC) care, placements & kinship, LADO (allegations against staff), MARAC/MAPPA interfaces, Missing/Exploited (CSE/CCE), education/health partners, police, courts/Cafcass, and information-sharing.

An *event* is a child/family outcome with adjudication potential (referral/triage, thresholding, safety plan, investigation, placement, contact, education/health provision, legal action, privacy). $R_t \in \{0,1\}$: **rejection** (culpable rights/safety breach or harm $\geq h_{\min}$); $B_t \in \{0,1\}$: breach (law/code/standard: statutory timescales, Working Together, information-sharing, LADO, placement standards, privacy/data); $C_t \in \{0,1\}$: culpability gate (intent/reckless/gross negligence by LA/partner/provider/vendor); $H_t \geq 0$: realized harm (significant harm, injury, exploitation episodes, missing days, unlawful contact/restriction, placement breakdown harm, privacy harm). $M_t, J_t^y \in \{0,1\}$: enacted mercy/justice (restoration to the *actually* harmed; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: increments to care/protection (love) and exercised rights/freedoms (freedom).

Illustrative incident classes $c \in C$: **TRIAGE-DEL** triage/threshold delay or error; **STRAT-FAIL** late/missed strategy meeting/s47; **CP-PLAN** CP plan not enacted (visits/safety actions missed); **PLAC-BRK** unsafe placement breakdown/unsuitable placement; **CSE-CCE** exploitation response failure (sexual/criminal); **MISS-RSP** missing episodes response failure; **LADO-FAIL** allegation against staff mishandled; **EDU-SAFE** school safeguarding failure (transfer, DSL escalation); **HEALTH-SAFE** health safeguarding/flag failure (maternity/HV/GP); **DATA-PRIV** privacy/data breach (case notes/photos/reports).

Switches: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0,1\}$ per class. Evaluator $\mathcal{J}$ is event-valued (Ch. 9).

36.1 Why an event-valued ledger fits safeguarding

Children experience *events*: were risks triaged in time; did a strategy meeting happen; was a safety plan delivered; did the placement keep them safe; did missing/exploitation stop; did breaches recur? Counting policies, training days, or partner MoUs over-credits *capacity*. We therefore score realized outcomes (H_t, R_t), enacted remedies (M_t, J_t^v), and whether hazardous routes are *structurally* and *provably* closed (HZ = 0, $S^* = 1$).

36.2 Operational semantics (instantiation)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \,(\text{rights/standard breach}) \,\text{or}\, H_t \geq h_{\min}\big) \,\text{and}\, C_t = 1.$$

$C_t = 1$ when an accountable LA/partner/provider/vendor had adequate knowledge/freedom and crossed a fault threshold (intent, recklessness, gross negligence) against a published *safety/ethics case*. If $C_t = 0$, log H_t for learning but do not set $R_t = 1$.

Harm H_t. Significant harm as per statute; injury; exploitation episodes (CSE/CCE); missing days/nights; unsafe contact; placement breakdown with harm; unlawful restriction; privacy harms. Publish $h_{\min}$ and class-specific harm markers.

Mercy/Justice M_t, J_t^v. $M_t = 1$ when *delivered*: immediate safety actions, placement move/support, therapy/health input, contact corrections, apology/compensation, record correction. $J_t^v = 1$ when *rectification/accountability* is enacted: process/tool change, staffing floors, partner escalation, vendor/system patch, sanctions/regulator notice.

Near-misses & drills. Near-miss (late but safe strategy; school DSL catches risk pre-harm) is not $R_t = 1$; convert to *eventized drills* ($D_t = 1$): MASH triage sims, s47 decision drills, missing response table-tops, exploitation disruption red-teams. Drills support S^* but do not mint M, J^v.

36.3 Evaluator and constraints (safeguarding view; broken to fit)

$$\mathcal{J}_{\text{cp}} = \sum_{t\geq 0}\Big(\alpha\,\Delta L_t + \gamma\,\Delta F_t + \mu\,M_t + \nu\,J_t^v\Big) - \sum_{t\geq 0}\Big(\beta\,H_t + \kappa\,R_t\Big), \tag{36.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau\leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^* \in \{0,1\}, \text{ updates per Ch. 8.}$$

Interpretation: ΔL_t credits *delivered* protections (timely visits, enacted plans, safe placements, disruption of exploitation, missing return interviews); ΔF_t credits *exercised* rights (lawful contact, advocacy/voice, education/health access, privacy).

36.4 Minimal-trigger doctrine (LA/partnership policy)

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (triage thresholds, visit cadence, plan delivery, placement sourcing, missing/exploitation response, LADO, information-sharing, vendor tools), fix, tests, owner+deadline, privacy-preserving publication plan.

2. **Remedy:** Enact M_t (safety actions, placement/kinship support, therapy/health, contact fixes, apology/compensation, record correction) and J_t^y (rectification/accountability: protocol/tool change, staffing floor, partner MoU escalations, sanctions/regulator notice).

3. **Confirm:** Stronger-incentive reenactment: cohort replay $\rightarrow$ adversarial/stress drills (out-of-hours, school transfers, missing spikes) $\rightarrow$ monitored live audits. On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats add cost (κ and H_t) but no new kind-level credit (typal/concave). One-off + closure dominates (Ch. 11).

36.5 Incident classes & scope locks (examples)

- **TRIAGE-DEL** — MASH/Front Door delay or thresholding error causing harm; referrals not screened/escalated in time.
- **STRAT-FAIL** — strategy meeting/s47 not held to timescales or without key partners; risk unmanaged.
- **CP-PLAN** — CP plan not delivered (visits missed; actions not completed) leading to harm/recurrence.
- **PLAC-BRK** — placement selection/safeguards fail; breakdown with harm; unsuitable B&B/URA use.
- **CSE-CCE** — exploitation response failure (intel sharing, disruption, safe exit) with continued episodes.
- **MISS-RSP** — missing episodes response failure (no return interview; patterns not disrupted).
- **LADO-FAIL** — staff/volunteer allegation mishandled (late/unsafe decisions).
- **EDU-SAFE** — school safeguarding transfer/DSL escalation failure; off-rolling/illegal exclusions with harm.

- **HEALTH-SAFE** — maternity/HV/GP safeguarding failures (flags not acted; info not shared).

- **DATA-PRIV** — privacy/data breach (case notes/photos unredacted; unsafe portals/sharing).

Each class has a written *scope lock* (inputs, thresholds, tools, preconditions); new causal vectors open c'.

36.6 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Significant harm; injury; exploitation episodes; missing days/nights; unsafe contact; breakdown harm; privacy harms
B_t (breach)	Statutory timescale breaches; Working Together/code violations; LADO mishandling; placement standard fails; information-sharing/privacy breaches
C_t (culpability)	Safety-case gap vs. standard; foreseeability; MASH/visit/plan logs; partner minutes; placement checks; LADO records; vendor audit logs
M_t (mercy)	Safety plan enacted; placement/kinship support; therapy/health input; contact corrections; apology/compensation; record correction
J_t^v (justice)	Protocol/tool change shipped; staffing floors; partner MoU escalations; sanctions/discipline; regulator/ombudsman notices
ΔL_t	Delivered protections: visits on cadence; plan actions done; safe placement sustained; missing return interviews; disruption activity
ΔF_t	Rights exercised: advocacy/voice; education/health access; lawful contact; privacy/data rights fulfilled

36.7 Confirmation tests (design pattern for LA partnerships)

Step 1: Cohort replay. Re-run matched cohorts (age/need/risk, partner mix, out-of-hours proportion) showing removal of the prior vector; use SPC or matched controls.
Step 2: Adversarial/stress drills. Out-of-hours triage; school transfer spikes; missing surges; exploitation disruption exercises; LADO response drills; pre-register pass metrics.
Step 3: Monitored live. Short live window with independent observers (IROs/LSCP/inspectorate), random case/placement audits, missing/exploitation dashboards. On pass, set $S^*(c) = 1$.
Scope lock. Freeze class scope; later failures with different causality open c'.

36.8 Tool-gating and interim safeguards after a trigger

Until $S^*(c) = 1$: enforce visit cadence floors; immediate safety actions; out-of-hours duty escalation; placement gating (no B&B for under-18s where prohibited); mandatory return interviews; exploitation disruption orders; LADO timelines with senior sign-off; privacy redaction defaults; vendor change-control freezes. Remove gates only after closure.

36.9 Micro-vignettes (worked examples)

(A) Late strategy/s47 (STRAT-FAIL). *Trigger:* Strategy delayed; injury occurs; culpability confirmed $\Rightarrow R_t = 1$.
Remedy: Safety actions; apology; plan cadence redesign; partner escalation protocol; vendor alerting ($M_t, J_t^y = 1$).
Confirm: Cohort replay; out-of-hours drills; monitored live audits; $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Exploitation continues (CSE-CCE). *Trigger:* Intel not shared; disruption not enacted; repeated episodes $\Rightarrow R_t = 1$.
Remedy: Disruption orders; placement/security changes; therapy; family support; multi-agency hub fix ($M_t, J_t^y = 1$).
Confirm: Red-team offender tactics; stress drills with nighttime spikes; monitored live; closure on pass.

(C) Placement breakdown with harm (PLAC-BRK). *Trigger:* Unsafe placement choice/monitoring; breakdown; missing episodes $\Rightarrow R_t = 1$.
Remedy: Move/support; kinship wrap; strengthened checks; provider sanctions ($M_t, J_t^y = 1$).
Confirm: Matched placement replay; provider stress tests; unannounced visits; class closed on pass.

36.10 Dashboards and metrics (ledger view)

- **Harm/recurrence:** significant harm; exploitation episodes; missing days; placement breakdowns; recurrence by class;

time-to-remedy; T_c^*.

- **Closure:** # classes with $S^* = 1$; median T_c^*; age/backlog of open dockets.

- **Culture (love):** visit cadence met; plan actions completed; return interviews done; disruption actions recorded (ΔL_t proxies).

- **Freedom:** advocacy/voice recorded; lawful contact; education/health access; privacy/data rights fulfilled (ΔF_t).

36.11 Anti-gaming and integrity

- **Under-recording.** Cross-check school/health/police logs; protected disclosures; random audits; publish privacy-preserving summaries.

- **Metric laundering.** Publish distributions (visit lateness, missing durations) not just means; pre-register pass criteria.

- **Scope creep.** Lock class definitions; do not relabel repeats as "new" unless causal vector differs.

- **Equity.** Confirm closure holds across age/sex/ethnicity/needs; report subgroup metrics in confirmation packs.

36.12 One-page checklist (drop-in for LA partnerships)

Safeguarding Incident → Confirmation Checklist

Trigger captured? MASH/visit/plan logs, partner minutes, placement checks, LADO records, dashboards
Gate set? visit cadence floors; immediate safety actions; placement gating; missing return interviews; LADO timelines
Docket opened? root cause, fix, tests, owner & deadline, publish plan
Remedy enacted? safety/placement/therapy/supports; contact fixes; apology/compensation; accountability
Tests passed? cohort replay ✓ adversarial/stress drills ✓ monitored live audits ✓
Scope locked? class frozen; distinct vectors filed as c'
Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined
Gates removed? only post-closure; rationale logged

36.13 Limits and open problems

- **Attribution.** Multi-actor causality (LA, police, health, schools, providers, courts, vendors) complicates C_t; adopt shared-fault taxonomies.
- **Hidden harms.** Exploitation/online grooming under-

detected; pair quantitative KPIs with qualitative panels and child voice.

- **Resource constraints.** Placement scarcity and caseloads confound fixes; use T_c^* targets, harm-based triage, and transparency.

36.14 What this chapter contributes

A full translation of the minimal-trigger doctrine to child protection & safeguarding: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and evidence standards. With typal/concave credits and $\kappa > 0$, repeating the same vector adds cost without adding kind-level value; one-off and confirmed closure dominates. “‘□0■

Part VI

Civic & Critical Systems

Chapter 37

Elections & Civic Integrity: Events, Remedies, and Confirmation

Notation

Domain mapping: voter registration, polling ops, accessibility, ballot issue/return, tabulation, audits/recounts, results publication, legal challenge, platform integrity, privacy.
$R_t \in \{0, 1\}$: **rejection** (culpable rights/process breach or harm $\geq h_{\min}$); B_t: breach (statute/procedure, accessibility, chain-of-custody, tabulation, privacy); C_t: culpability (intent/reckless/gross negligence); $H_t \geq 0$: realized harm (disenfranchisement events, spoiled ballots due to official error, custody breaks, miscount, privacy exposure).

$M_t, J_t^v \in \{0,1\}$: enacted mercy/justice (restoration to those actually harmed; rectification/accountability); $\Delta L_t, \Delta F_t \geq 0$: increments to civic care (service delivered) and exercised rights/freedoms (votes cast & counted).
Incident classes $c \in C$: **REG-ERR**, **ACCESS-FAIL**, **BALLOT-ISSUE**, **CHAIN-CUST**, **TAB-ERR**, **AUDIT-FAIL**, **PRIV-LEAK**, **PLATFORM-CIVIC**.
Switches: $\mathrm{HZ}_c, S_c^* \in \{0,1\}$. Evaluator $\mathcal{J}$ as in Ch. 9.

37.1 Why events decide legitimacy

Legitimacy rests on *events*: eligible voters actually cast ballots, ballots are actually counted correctly, disputes are actually resolved. Policies and manuals are *capacities*; our ledger credits realized service and penalizes realized failures, then requires structural closure ($S^* = 1$, $\mathrm{HZ} = 0$).

37.2 Operational semantics

Rejection R_t.

$$R_t = 1 \iff \big(B_t = 1 \,(\text{procedure/rights breach}) \,\text{or}\, H_t \geq h_{\min}\big) \,\text{and}\, C_t = 1.$$

Examples: wrong ballot style issued, chain-of-custody lapse, tabulation misconfiguration, inaccessible polling place, privacy breach.

Mercy/Justice. $M_t = 1$ when harmed voters are restored (re-issue ballots, extended hours, corrected count, privacy remediation). $J_t^v = 1$ for rectification/accountability (procedure/tool change, sanctions, public audit logs).

Near-misses & drills. Near-miss (L&A caught an error) is not $R_t = 1$; convert to eventized drills ($D_t = 1$): adversarial L&A, custody exercises, mock audits, platform civic-policy red-teams. Drills support S^* but do not mint M, J^v.

37.3 Evaluator and constraints

$$\mathcal{J}_{\text{civic}} = \sum_{t \geq 0} \left(\alpha \Delta L_t + \gamma \Delta F_t + \mu M_t + \nu J_t^v \right) - \sum_{t \geq 0} \left(\beta H_t + \kappa R_t \right), \tag{37.1}$$

$$\text{s.t. } E_{\text{tot}} \leq H_{\max}, \quad \text{HZ}_c, S_c^* \in \{0, 1\}. \tag{37.2}$$

37.4 Minimal-trigger doctrine

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ remediation docket (root cause, fix, tests, owner, deadline, publish plan).
2. **Remedy:** M_t (restore voters/results; privacy fixes), J_t^v (procedure/tool change; sanctions; public audit logs).

3. **Confirm:** L&A with adversarial decks; custody stress runs; mock risk-limiting audit; monitored live checks. On pass, $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats add costs but no new kind-level credit; one-off + closure dominates.

37.5 Evidence standards

Indicator	Event-valued evidence
H_t (harm)	documented disenfranchisement; miscount corrected; privacy exposure; accessibility denial events
B_t (breach)	statute/procedure violations; custody gaps; tabulation config errors; audit not executed; privacy breaches
C_t (culpability)	logs/configs/custody forms/audit trails; foreseeability vs standards
M_t (mercy)	ballot re-issue; hours extended; corrected results; privacy remedies; notices
J_t^v (justice)	tool/procedure change; accountability/sanctions; public audit logs
ΔL_t	polling open/accessible; correct ballots issued; audits complete; publication SLAs met
ΔF_t	votes cast and counted; corrections applied; legal rights exercised
S_c^*	passed adversarial L&A + monitored live audit; scope locked

37.6 Micro-vignettes

(A) Wrong ballot style. Trigger $\Rightarrow R_t = 1$; Remedy: re-issue, notify, adjust counts (M, J^v); Confirm: adversarial L&A; monitored live.

(B) Chain-of-custody gap. Trigger via audit; Remedy: redesign custody/seals; Confirm: stress runs, surprise audits; $S^* = 1$.

(C) Tabulation misconfig. Trigger via miscount; Remedy: recount & config hardening; Confirm: red-team decks; monitored live.

37.7 Checklist

Election Incident → Confirmation Checklist

Trigger captured? Gate set? Docket opened? Remedy enacted? Tests passed? Scope locked? Closure set?

37.8 Limits

Attribution across officials/vendors; rare-but-severe failures; privacy constraints. Use shared-fault taxonomies, periodic re-confirmations for high-risk classes, and privacy-preserving publication.

37.9 Contribution

Translates minimal-trigger doctrine to elections: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), evidence standards; typal credits + $\kappa > 0$ make repetition strictly dominated.

Chapter 38

Disaster & Emergency Management: Incidents, Command Integrity, and Confirmation

Notation

Domain mapping (multi-agency): risk assessment, public warning/alerting, evacuation/shelter, firefighting/flood response, urban search & rescue, EMS/mass-casualty (MCI), hospitals/bed surge, public health (epidemics/IPC), utilities continuity, logistics & supply, critical comms/IT, public information, recovery. ICS/Gold-Silver-Bronze command.

Event: a person/community/system outcome with adjudication potential (e.g., failure to warn, late/unsafe evacuation, shelter not opened, triage misclassification, hospital diversion collapse, contamination exposure, comms blackout, logistics shortfall). Variables per time t and incident class $c \in C$:

- $R_t \in \{0, 1\}$: **rejection** = culpable rights/safety breach or harm $\geq h_{\min}$.
- $B_t \in \{0, 1\}$: breach (law/code/standard/SOP: ICS doctrine, evacuation/shelter standards, MCI/triage, water/air quality limits, public-info protocols, continuity obligations).
- $C_t \in \{0, 1\}$: culpability gate (intent/reckless/gross negligence by agency/provider/vendor).
- $H_t \geq 0$: realized harm (fatalities/injuries, exposure days, unserved evacuations, shelterless nights, service-loss minutes, triage errors).
- $M_t, J_t^v \in \{0, 1\}$: enacted mercy/justice (restoration/accountability actually delivered).
- $\Delta L_t, \Delta F_t \geq 0$: increments in protection (love) and exercised freedoms/rights (freedom) *actually delivered* (alerts issued, accessible evacuation, shelter staffed, care provided).

Illustrative incident classes C: **WARN-FAIL** (public alerting failure); **EVAC-FAIL** (late/unsafe evacuation); **SHELTER-FAIL** (shelter/ACC not opened/accessible); **FIREBREAK-FAIL** (containment/line breach); **FLOOD-DEF** (barrier/pump failure); **MCI-TRIAGE** (mass-casualty triage/transport breakdown); **HOSP-SURGE** (ED diversion/bed surge collapse); **PANDEM-IPC** (infection prevention/control breach); **COMMS-INFO** (public information error/contradiction); **LOG-RESUP** (logistics/resupply failure); **WATER-AIR** (quality exceedance); **CYBER-EMIS** (control room/dispatch/911 outage).
Switches per class: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0, 1\}$. Evaluator $\mathcal{J}$ is event-valued (§ 9).

38.1 Why an event-valued ledger fits incidents

In disasters, only *events* save lives: did warnings actually reach people, did evacuation actually occur in time, did shelters actually open, did triage actually route patients correctly, did water remain safe. Policies, binders, and exercises are *capacity*; we credit realized protections ($\Delta L_t, \Delta F_t$), debit realized costs (H_t) and culpable breaches (R_t), and require structural closure ($S^* = 1, \mathrm{HZ} = 0$) before removing interim gates.

38.2 Operational semantics (instantiation)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \,(\text{rights/standard breach}) \,\text{or}\, H_t \geq h_{\min}\big) \,\text{and}\, C_t = 1.$$

$C_t = 1$ when an accountable agency/provider/vendor had adequate knowledge/freedom and crossed a fault threshold versus a published *safety/ethics case*. If $C_t = 0$ (accident), log H_t for learning but do not set $R_t = 1$.

Evaluator (EM view).

$$\mathcal{J}_{\text{em}} = \sum_{t \geq 0} \Big(\alpha\, \Delta L_t + \gamma\, \Delta F_t + \mu\, M_t + \nu\, J_t^{v} \Big) \;-\; \sum_{t \geq 0} \Big(\beta\, H_t + \kappa\, R_t \Big), \tag{38.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau \leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^{*} \in \{0, 1\} \text{ per class.}$$

Near-misses & drills. Near-miss (sirens late but neighborhood self-evacuates; fireline holds due to luck) is not $R_t = 1$. Convert to *eventized drills* ($D_t = 1$): full-stack alert tests (cell broadcast/sirens), evacuation table-tops + bus convoys, shelter activation with real staff/stock, MCI triage simulations, pump/valve functional tests, control-room failovers, cyber red-teams. Drills support S^{*} but do not mint M, J^{v}.

38.3 Minimal-trigger doctrine (multi-agency policy)

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (doctrine, staffing, tooling, logistics, vendor, interop), fix, tests, owner+deadline, privacy-preserving publication.
2. **Remedy:** $M_t = 1$ (*delivered* restoration/aid/compensation/record correction) & $J_t^v = 1$ (rectification/accountability: tool-gates, doctrine changes, sanctions/vendor fixes).
3. **Confirm:** Stronger-incentive reenactment: cohort replay $\rightarrow$ stress drills (night/weekend, high wind/flow, comms outage) $\rightarrow$ monitored live checks. On pass, set $S^*(c) = 1, \mathrm{HZ}(c) = 0$.
4. **No repetition bonus:** Repeats add cost (κ and H_t) but no new kind-level credit (typal/concave). One-off + closure dominates.

38.4 Incident classes & scope locks (examples)

- **WARN-FAIL** — public alert not issued/propagated/accessible (languages, disability) in time.

- **EVAC-FAIL** — evacuation order late/contradictory; transport not provided; blocked routes.
- **SHELTER-FAIL** — shelter/ACC not opened, unstaffed, inaccessible, unstocked (heatwave/cold).
- **FIREBREAK-FAIL** — containment line breaks; backburns mis-timed; comms/air ops coordination failure.
- **FLOOD-DEF** — barrier/pump failure; gate mis-ops; levee breach.
- **MCI-TRIAGE** — triage/transport breakdown; wrong hospital loads; ambulance dispatch outage.
- **HOSP-SURGE** — ED diversion cascade; bed surge planning fails; oxygen/ICU/utilities lapses.
- **PANDEM-IPC** — IPC breach; PPE rationing failure; cohorting/ventilation lapses.
- **COMMS-INFO** — wrong/contradictory public info; no rumor control; languages omitted.
- **LOG-RESUP** — fuel/food/water/medical supply failure; unmapped dependencies.
- **WATER-AIR** — drinking-water/air quality exceedance without timely remedy.
- **CYBER-EMIS** — control room/dispatch/911/999 outage; EOC software failure.

Each class has a written *scope lock* (inputs, thresholds, tools, preconditions). New causal vectors open c'.

38.5 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	Fatalities/injuries; exposure days; unserved evacuation requests; shelterless nights; service-loss minutes; triage misclassification/error rates
B_t (breach)	ICS/SOP non-compliance; evacuation/shelter standards breached; MCI/triage protocol failures; quality exceedances; public-info/continuity obligations breached
C_t (culpability)	Logs (EOC/dispatch), ICS forms, call records, GIS/AVL, timestamps, staffing rosters, vendor tickets; foreseeability vs standards
M_t (mercy)	Delivered aid/compensation/temporary housing; rehydration/cooling/heating; corrected public records; replacement IDs; health follow-up
J_t^v (justice)	Doctrine/tool changes shipped; sanctions/vendor remedies; regulator notices; published closure pack
ΔL_t	Alerts issued; evac routes cleared; shelters staffed/stocked; pumps/valves operated; triage correct; hospital beds opened; IPC implemented
ΔF_t	Accessible warnings; accessible transport; right to shelter/aid exer-

38.6 Confirmation tests (design pattern)

Step 1: Cohort replay. Re-run matched hazards (season, geography, population vulnerability) showing removal of the prior vector; include matched controls/SPC.
Step 2: Stress drills. Night/weekend staffing; high winds/flow; simultaneous incidents; comms/cyber failovers; hospital surge; supply shocks; pre-register pass metrics.
Step 3: Monitored live. Short live window with independent observers (mutual aid/inspectorate), random checks (sirens, shelters, pumps), dispatch/hospital telemetry. On pass, set $S^*(c) = 1$.
Scope lock. Freeze class scope; later failures with different causality open c'.

38.7 Tool-gating & interim safeguards after a trigger

Until $S^*(c) = 1$: *alert-must-run* rules; multilingual/accessible comms; evacuation bus convoys; contraflow approvals; shelter staffing floors; generator/fuel stock floors; hospital diversion caps; IPC hard-stops (PPE, zoning); dispatch/control-room fallbacks; vendor change-control freeze. Remove gates only after closure.

38.8 Micro-vignettes (worked examples)

(A) Late evacuation messaging (EVAC-FAIL). *Trigger:* Mixed messages; trapped households; injuries/fatalities; culpability con-

firmed $\Rightarrow R_t = 1$.
Remedy: Unified messaging doctrine; multi-channel alerts; route clearance; transport convoys; aid/compensation ($M_t, J_t^v = 1$).
Confirm: Cohort replay; night/weekend drills; monitored live siren/cell tests; $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Wildfire line breach (FIREBREAK-FAIL). *Trigger:* Comms/air-ground coordination failure; line lost; property loss; injuries $\Rightarrow R_t = 1$.
Remedy: Air-ground comms redesign; checklists; weather windows; staging; vendor radio patch ($M_t, J_t^v = 1$).
Confirm: Adversarial burn-over drills; wind-stress exercises; monitored live telemetry; closure on pass.

(C) MCI triage breakdown (MCI-TRIAGE/HOSP-SURGE). *Trigger:* Triage misclassifications; ED gridlock; diversions fail; harm occurs $\Rightarrow R_t = 1$.
Remedy: Triage retrain with device prompts; load-balancing; surge beds; EMS/hospital comms SOP; compensation where due ($M_t, J_t^v = 1$).
Confirm: Full-scale MCI drill; EMS-to-ED stress tests; monitored live dashboards; set $S^* = 1$.

38.9 Dashboards & metrics (ledger view)

- **Harm/recurrence:** fatalities/injuries; exposure days; unserved evac/shelter requests; service-loss minutes; recurrence

by class; time-to-remedy; T_c^*.

- **Closure:** # classes with $S^* = 1$; median T_c^*; backlog/age of open dockets.

- **Culture (love):** alert reach/time; shelters opened/staffed; pumps running; hospital surge beds; IPC adherence (ΔL_t).

- **Freedom:** accessible alerts/transport/shelter; lawful movement; data/privacy in relief (ΔF_t).

38.10 Anti-gaming and integrity

- **Paper compliance.** No credit for plans alone; require event artefacts (timestamps, logs, telemetry).

- **Metric laundering.** Publish distributions (alert latency, evac arrival times) not just means; pre-register pass criteria.

- **Under-recording.** Protected disclosures; independent observers; random checks; privacy-preserving public summaries.

- **Equity.** Confirm closure holds across languages, disability, age, and neighborhoods; include subgroup metrics.

38.11 One-page checklist (drop-in for EOCs)

Incident → Confirmation Checklist (EOC/Gold)

Trigger captured? alerts/dispatch/EOC logs, ICS forms, GIS, hospital dashboards

Gate set? alert-must-run; accessible comms; evac convoys; shelter staffing floors; surge/IPC/tool gates

Docket opened? root cause, fix, tests, owner & deadline, publish plan

Remedy enacted? aid/compensation; record corrections; doctrine/tool change; vendor sanctions

Tests passed? cohort replay ✓ stress drills (night/weather/multi-incident/comms failover) ✓ monitored live checks ✓

Scope locked? class frozen; distinct vectors filed as c'

Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined

Gates removed? only post-closure; rationale logged

38.12 Limits and open problems

- **Attribution** across agencies/vendors complicates C_t; adopt shared-fault taxonomies.

- **Rare but severe** events require reliance on drills/inspection; preserve honesty locks to avoid "pretend passes."

- **Privacy/ethics** in publishing artefacts; use privacy-preserving summaries and redaction standards.

38.13 What this chapter contributes

A complete translation of the minimal-trigger doctrine to disaster & emergency management: precise triggers (R_t), docketed remedies (M, J^{ν}), structural closure (S^{*}), and event-valued evidence standards. With typal/concave credits and $\kappa > 0$, repetition of the same vector adds cost without kind-level value; one-off + confirmed closure dominates while preserving live freedom to act. “‘□0■

Chapter 39

Finance & Banking: Controls, AML/Conduct, and Confirmation

Notation

Domain mapping (retail & wholesale): onboarding/KYC/CDD, sanctions/PEP screening, transaction monitoring, fraud prevention (card, account takeover, APP scams), payments (wires/ACH/SEPA/RTGS), trading/market abuse, suitability & mis-selling, complaints & redress, credit underwriting & collections, operational resilience (payments/core), data privacy & security, vendors/fintech partners, reporting to supervisors.

Event: a customer/market/system outcome with adjudication potential: unlawful transaction allowed, legitimate payment wrongly blocked, sanctions hit processed, fraud loss, client harmed by mis-selling, market manipulation undetected, privacy breach, outage preventing access to funds.
Variables per time t and incident class $c \in C$:

- $R_t \in \{0, 1\}$: **rejection** = culpable rights/safety/market-integrity breach or realized harm $\geq h_{\min}$.
- $B_t \in \{0, 1\}$: breach (law/code/standard: AML/CFT, sanctions, conduct/suitability, market abuse, operational resilience, data protection).
- $C_t \in \{0, 1\}$: culpability gate (intent/reckless/gross negligence by firm/desk/ops/vendor).
- $H_t \geq 0$: realized harm (customer/victim loss, unlawful funds flow, regulatory penalty, market distortion, denial-of-service to legitimate users, privacy harms).
- $M_t, J_t^v \in \{0, 1\}$: enacted mercy/justice (restoration to those actually harmed; rectification/accountability).
- $\Delta L_t, \Delta F_t \geq 0$: increments to protection (love)—customer protection, market integrity— and to exercised lawful freedoms/rights (freedom)—ability to access and move funds, trade, privacy rights—*actually delivered*.

Illustrative incident classes C: **KYC-CDD** onboarding due diligence failure; **SANCTIONS** sanctions/PEP screening miss; **STR-LATE** late/nonfiled suspicious report; **WIRE-FRAUD** authorised push payment (APP) fraud miss; **CARD-FRAUD** card-not-present takeover miss; **AUTHZ-FAIL** wrongful blocks/denials to legitimate users; **MKT-ABUSE** market manipulation/insider abuse miss; **MIS-SELL** suitability/mis-selling harm; **OPS-OUTAGE** payments/core outage (access to funds); **MODEL-DRIFT** monitoring model drift causing harms; **DATA-PRIV** privacy/data breach; **CONF-LIB** conflicts/Chinese wall failure.
Switches per class: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0,1\}$. Evaluator $\mathcal{J}$ is event-valued (cf. formal core).

39.1 Why an event-valued ledger fits finance

Financial systems are judged by *events* that happen to real customers and markets: Did illicit funds actually pass? Were legitimate customers actually blocked from rent/medicine? Were victims *restored*? Did the same vector recur? Policies, model docs, and "strong intent" are *capacities*; we score realized protections $(\Delta L_t, \Delta F_t)$, subtract realized costs (H_t), count culpable rejections (R_t), and require structural closure ($S^* = 1$, $\mathrm{HZ} = 0$) before removing temporary gates.

39.2 Operational semantics (instantiation)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \text{ (law/standard breach) or } H_t \geq h_{\min}\big) \text{ and } C_t = 1.$$

$C_t = 1$ where an accountable team/vendor had adequate knowledge/freedom and crossed a fault threshold relative to a published *safety/ethics case* (controls, limits, monitoring, and playbooks). If $C_t = 0$ (accident), log H_t for learning but do not set $R_t = 1$.

Two-sided harm. Missed bad activity and overblocking legitimate users both harm:

$$H_t \;=\; H_t^{\text{miss}} \;+\; \lambda\, H_t^{\text{over}}, \qquad \lambda > 0,$$

with λ set via governance to reflect the seriousness of lawful denials (e.g., rent/medicine payments).

Evaluator (finance view).

$$\mathcal{J}_{\text{fin}} = \sum_{t\geq 0} \Big(\alpha\, \Delta L_t + \gamma\, \Delta F_t + \mu\, M_t + \nu\, J_t^{v}\Big) \;-\; \sum_{t\geq 0} \Big(\beta\, H_t + \kappa\, R_t\Big), \tag{39.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau\leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^* \in \{0,1\} \text{ per class.}$$

Interpretation: ΔL_t credits *delivered* protections (fraud actually prevented with restitution, illicit funds actually blocked, market manipulation actually disrupted). ΔF_t credits *exercised* rights (lawful

access to funds, fair trading, privacy exercised).

Near-misses & drills. Near-miss (alert caught pre-settlement; false-positive reversed before harm) is not $R_t = 1$. Convert to *eventized drills* ($D_t = 1$): adversarial payments (synthetic mule rings), sanctions fuzzing, red-team insider scenarios, latency/load failovers, model shadow tests. Drills support S^* but do not mint M, J^v.

39.3 Minimal-trigger doctrine (firm policy)

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (control gap, model drift, thresholds, latency, vendor, staffing), fix, tests, owner+deadline, privacy-preserving publication plan.

2. **Remedy:** $M_t = 1$ (*delivered* restitution/chargeback/fee reversal/credit repair/record correction) and $J_t^v = 1$ (rectification/accountability: tool-gates, model re-train, sanctions list sync, desk controls, sanctions/vendor remedies).

3. **Confirm:** Stronger-incentive reenactment: cohort replay $\rightarrow$ adversarial probes (synthetic fraud rings, sanctions fuzzing, insider patterns) $\rightarrow$ monitored live. On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats add cost (κ and H_t) but *no* kind-level credit (typal/concave). One-off + closure dominates.

39.4 Incident classes & scope locks (examples)

- **KYC-CDD** — beneficial ownership missed; high-risk customer onboarded without EDD; synthetic ID pass.
- **SANCTIONS** — sanctioned party processed; list sync latency; fuzzy-match miss.
- **STR-LATE** — suspicious matter detected but report late/not filed.
- **WIRE-FRAUD** — APP scam unblocked; mule network allowed; recovery failed.
- **CARD-FRAUD** — takeover/CNP fraud undetected; alerts too slow.
- **AUTHZ-FAIL** — wrongful denials/holds (rent/medicine); discriminatory overblocking.
- **MKT-ABUSE** — spoofing/layering/insider trading undetected; surveillance failure.
- **MIS-SELL** — unsuitable product sold; disclosure failures; harm realized.

- **OPS-OUTAGE** — payments/core outage; access to funds blocked.

- **MODEL-DRIFT** — monitoring model drifts (recall/precision collapse); harm ensues.

- **DATA-PRIV** — privacy breach (PII/transactions); unlawful profiling/disclosure.

- **CONF-LIB** — conflict of interest; wall breach; misuse of MNPI.

Each class has a written *scope lock* (signals, thresholds, SLAs, tooling, preconditions). New causal vectors open c'.

39.5 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	customer/victim financial loss; illicit funds value; regulatory penalties; market impact; denial-of-service counts/durations; privacy harm
B_t (breach)	AML/sanctions/conduct/market abuse/privacy/resilience breaches; SLA/latency breaches tied to harm
C_t (culpability)	control inventory gap vs standard; thresholds/latency; model monitoring; change-control logs; surveillance/alert queues
M_t (mercy)	restitution/chargeback/rebate; fee reversal; credit repair; corrected records; public notices where applicable
J_t^v (justice)	model/tool changes shipped; thresholds/latency fixes; sanctions list/process fixes; desk controls; vendor sanctions; regulator notices
ΔL_t	fraud prevented; illicit flow blocked; market manipulation disrupted; outage mitigations
ΔF_t	lawful access to funds restored; fair trading; privacy/data rights fulfilled
S_c^*	passed cohort replay + adversarial probes + monitored live; artefacts linked (privacy-preserving)

39.6 Confirmation tests (design pattern)

Step 1: Cohort replay. Re-run matched periods/segments (seasonality, channel, geography, product) showing removal of the prior vector; use SPC or matched controls.
Step 2: Adversarial probes. Synthetic mule rings; sanctions fuzzing; insider/trade-pattern red-team; latency/failover exercises; pre-register pass metrics (recall/precision, false-positive fairness, loss stopped).
Step 3: Monitored live. Short live window with independent risk/audit/regulatory observers; random alert reviews; payment/trade sampling; fairness slices (age/sex/ethnicity where lawful). On pass, set $S^*(c) = 1$.
Scope lock. Freeze class scope; later failures with different causality open c'.

39.7 Tool-gating & interim safeguards after a trigger

Until $S^*(c) = 1$: raise frictions/limits on affected corridors; dual-approval on large/risky flows; sanctions "block on fuzzy-high"; enhance post-transaction recovery playbooks; surveillance priority queues; vendor change-control freeze; payments outage fallback (branch/cash/alt rails). Remove gates only after closure.

39.8 Micro-vignettes (worked examples)

(A) APP fraud ring (WIRE-FRAUD). *Trigger:* Multiple victims; mule network; recovery fails; culpability confirmed $\Rightarrow R_t = 1$.
Remedy: Restitution/chargebacks; corridor limits; behavioral model uplift; takedown with partner banks; vendor sanctions $(M_t, J_t^v = 1)$.
Confirm: Synthetic ring probes; weekend/night spikes; monitored live recall/precision; $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Sanctions miss (SANCTIONS). *Trigger:* Listed entity paid; list sync lag; fuzzy miss $\Rightarrow R_t = 1$.
Remedy: List sync SLOs; fuzzy tuning; manual tier for high-suspects; vendor penalty; beneficiary clawback where lawful $(M_t, J_t^v = 1)$.
Confirm: Fuzzed test decks; live shadow scans; pass thresholds; closure on pass.

(C) Mis-selling harm (MIS-SELL). *Trigger:* Unsuitable product sold; losses realized; complaints upheld $\Rightarrow R_t = 1$.
Remedy: Redress; disclosure redesign; suitability engine; staff incentives reset; records corrected $(M_t, J_t^v = 1)$.
Confirm: Adversarial suitability cases; monitored live complaints trend; class closed on pass.

39.9 Dashboards & metrics (ledger view)

- **Harm/recurrence:** victim loss; illicit flow value; denial-of-service counts/duration; regulatory fines; recurrence by class; time-to-remedy; T_c^*.

- **Closure:** # classes with $S^* = 1$; median T_c^*; backlog/age of open dockets.

- **Culture (love):** prevented fraud/illicit flows; redress delivered; complaints resolved (ΔL_t).

- **Freedom:** access-to-funds SLAs; false-positive reversal times; fair-surveillance slices; privacy rights fulfilled (ΔF_t).

39.10 Anti-gaming and integrity

- **Freeze-to-win.** Lowering loss by freezing legitimate customers is penalized via H_t^{over} and fairness slices.

- **SAR inflation.** Counting reports without substance earns no J_t^v; require event-change proof.

- **Latency laundering.** Report mean latency *and* distribution; pre-register pass criteria.

- **Scope creep.** Lock classes; do not relabel repeats as new vectors unless causality differs.

- **Equity.** Confirm closure holds across demographics/segments where lawful; publish subgroup metrics in closure packs.

39.11 One-page checklist (drop-in for first/second line)

Financial Incident → Confirmation Checklist

Trigger captured? alert logs, sanction scans, trade surveillance, payment traces, complaints

Gate set? corridor frictions/limits; dual-approvals; vendor change-freeze; outage fallbacks

Docket opened? root cause, fix, tests, owner & deadline, publish plan

Remedy enacted? restitution/redress; fee reversal; credit repair; accountability actions

Tests passed? cohort replay ✓ adversarial probes ✓ monitored live ✓

Scope locked? class frozen; distinct vectors filed as c'

Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined

Gates removed? only post-closure; rationale logged

39.12 Limits and open problems

- **Attribution** across firm/vendor/networks complicates C_t; adopt shared-fault taxonomies and interbank closure packs.
- **Rare, severe** vectors (nation-state sanctions evasion) rely on drills & intel; preserve honesty locks.

- **Privacy/competition** constraints on publishing artefacts; use privacy-preserving summaries and regulator-held confirmations.

39.13 What this chapter contributes

A full translation of the minimal-trigger doctrine to finance: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and event-valued evidence standards that balance *both* kinds of harm—missed bad activity and overblocking of the good. With typal/concave credits and $\kappa > 0$, repetition adds cost without kind-level value; one-off + confirmed closure dominates while protecting lawful freedom to transact. “‘□0■

Chapter 40

Supply Chain & Food Safety: Contamination, Recalls, and Confirmation

Notation

Domain mapping (farm → fork): primary production, feed, slaughter/processing, cold chain, manufacturing/packaging, storage & distribution, import controls, retail/food service, sanitation & hygiene, environmental monitoring, traceability (one-up/one-down), recalls & public comms, laboratory testing, vendor assurance.

Event: a consumer/worker/market/system outcome with adjudication potential (e.g., illness cluster linked to product, undeclared allergen exposure, cold-chain failure, foreign material, chem-

ical exceedance, traceability break, delayed/ineffective recall, sanitation lapse).
Variables per time t and incident class $c \in C$:

- $R_t \in \{0, 1\}$: **rejection** = culpable rights/safety breach or realized harm $\geq h_{\min}$.
- $B_t \in \{0, 1\}$: breach (law/code/standard: HACCP/FSMS, allergens, micro/chemical limits, GHP/SSOP, packaging integrity, traceability, recall obligations).
- $C_t \in \{0, 1\}$: culpability gate (intent/reckless/gross negligence by operator/transport/vendor/lab).
- $H_t \geq 0$: realized harm (illness/hospitalization/fatality counts; exposure days; product consumed before recall; economic loss to protected groups; food insecurity from over-broad recall; worker injury; privacy harm in case reporting).
- $M_t, J_t^y \in \{0, 1\}$: enacted mercy/justice (restoration to those actually harmed; rectification/accountability).
- $\Delta L_t, \Delta F_t \geq 0$: increments to protection (love) and exercised freedoms/rights (freedom) *actually delivered*: safe product released, targeted recall executed, consumer information/choice preserved.

Illustrative incident classes C: **PATHOGEN** (Listeria/Salmonella/E. coli contamination); **ALLERGEN** (un-

declared allergen); **TEMP-COLD** (cold-chain excursion); **FOREIGN-MAT** (metal/plastic/glass); **CHEM-RESID** (chemical residues/exceedances); **WATER-SAN** (wash/ice water lapses); **ENV-SWAB** (positive environmental swab not contained); **PACK-INTEG** (seal/leak/gas mix failure); **TRACE-FAIL** (traceability/lot mapping gap); **RECALL-SLOW** (late/ineffective recall); **VENDOR-SPEC** (supplier spec/audit failure); **LABEL-DATE** (mislabel/shelf-life error). Switches per class: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0,1\}$. Evaluator $\mathcal{J}$ is event-valued (cf. formal core).

40.1 Why an event-valued ledger fits food safety

Food safety succeeds or fails in *events*: did contaminated lots actually ship; did consumers actually ingest an undeclared allergen; did the recall actually reach households; did the same vector recur at the plant or haulier. Policies, HACCP binders, and vendor certificates are *capacity*; we credit realized protections ($\Delta L_t, \Delta F_t$), debit realized costs (H_t) and culpable breaches (R_t), and require structural closure ($S^* = 1$, HZ = 0) before removing interim gates.

40.2 Operational semantics (instantiation)

Rejection R_t (gate).

$$R_t = 1 \iff \big(B_t = 1 \text{ (law/standard breach) or } H_t \geq h_{\min}\big) \text{ and } C_t = 1.$$

$C_t = 1$ where an accountable operator/vendor/transport/lab had adequate knowledge/freedom and crossed a fault threshold relative to a published *safety/ethics case* (HACCP plan, monitoring, SSOPs, environmental program, traceability, recall playbook). If $C_t = 0$ (accident), log H_t for learning but do not set $R_t = 1$.

Two-sided harm (safety vs waste/hunger).

$$H_t \;=\; H_t^{\text{exposure}} \;+\; \lambda\, H_t^{\text{over-recall}}, \qquad \lambda > 0,$$

so over-broad recalls that needlessly waste safe food (or cause shortages for vulnerable groups) are penalized, pushing toward *targeted* recalls with strong evidence.

Evaluator (supply chain view).

$$\mathcal{J}_{\text{sc}} = \sum_{t\geq 0} \Big(\alpha\, \Delta L_t + \gamma\, \Delta F_t + \mu\, M_t + \nu\, J_t^{v}\Big) \;-\; \sum_{t\geq 0} \Big(\beta\, H_t + \kappa\, R_t\Big), \tag{40.1}$$

$$\text{s.t. } E_{\text{tot}} = \textstyle\sum_{\tau\leq t} H_\tau \leq H_{\max}, \quad \text{HZ}_c, S_c^* \in \{0,1\} \text{ per class.}$$

Interpretation: ΔL_t credits *delivered* protections (contamination contained, targeted recall, sanitation restored); ΔF_t credits *exercised* freedoms/rights (accurate labels, allergen info, fair access to safe product, privacy in case handling).

Near-misses & drills. Near-miss (presumptive positive caught pre-release; cold-chain data logger flags but product quarantined)

is not $R_t = 1$. Convert to *eventized drills* ($D_t = 1$): mock recalls, adversarial lot-mixing tests, cold-chain power-failure simulations, sanitation challenge tests, traceability sprints, lab ring trials. Drills support S^* but do not mint M, J^v.

40.3 Minimal-trigger doctrine (operator policy)

Per incident class c (e.g.

1. **Trigger:** First adjudicated $R_t = 1$ of class $c \Rightarrow$ open a *remediation docket*: root cause (HACCP hazard analysis, SSOPs, supplier control, sanitation design, equipment, packaging, transport, lab methods), fix, tests, owner+deadline, privacy-preserving publication.

2. **Remedy:** $M_t = 1$ (*delivered* consumer restitution/medical costs where applicable, product replacement/refund, public notice, record correction) and $J_t^v = 1$ (rectification/accountability: process/tool change, supplier sanctions, packaging redesign, lab/vendor remedies).

3. **Confirm:** Stronger-incentive reenactment: cohort replay $\rightarrow$ stress drills (peak throughput, hottest ambient, longest legs, power/CO_2 excursions, line restarts) $\rightarrow$ monitored live (environmental swabs, data loggers, targeted holds). On pass, set $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$.

4. **No repetition bonus:** Repeats add cost (κ and H_t) but no new kind-level credit (typal/concave). One-off + closure dominates.

40.4 Incident classes & scope locks (examples)

- **PATHOGEN** — product or environment positive (Listeria in RTE zone, Salmonella in raw, STEC in leafy greens).
- **ALLERGEN** — undeclared allergen from mislabel, changeover failure, rework contamination.
- **TEMP-COLD** — time/temperature abuse; logger gaps; door-open cycles; last-mile failure.
- **FOREIGN-MAT** — metal/plastic/glass from equipment/wear/packaging.
- **CHEM-RESID** — sanitizer/pesticide residues; migration from inks/plastics; nitrite/nitrate exceedance.
- **WATER-SAN** — process/ice water fails micro/chemical limits.
- **ENV-SWAB** — zone 2/3 Lm positives not contained; vector to product zone.
- **PACK-INTEG** — seal/gas mix failure; MAP leak; vacuum loss.

- **TRACE-FAIL** — lot mapping/one-up-one-down breaks; time to trace exceeds target.

- **RECALL-SLOW** — recall initiation/communication delay; low consumer reach.

- **VENDOR-SPEC** — supplier adulteration/mis-spec; audit failure.

- **LABEL-DATE** — wrong date/keep conditions; misleading claims.

Each class has a written *scope lock* (signals, thresholds, tools, preconditions). New causal vectors open c'.

40.5 Evidence standards (what counts as "event")

Indicator	Acceptable event-valued evidence
H_t (harm)	illness/hospitalization/fatality counts; exposure days; consumed units prior to recall; vulnerable-group impact; food insecurity/waste from over-recall
B_t (breach)	HACCP/FSMS/SSOP non-conformance; micro/chemical exceedances; allergen code breaches; traceability/recall obligations; packaging integrity failures
C_t (culpability)	HACCP plan gaps; sanitation/verification logs; environmental swabs; line changeover records; lab methods/CoA; transport logger data; foreseeability vs standards
M_t (mercy)	refunds/replacements; medical/recovery support; targeted public notices; record correction; disposal with equity mitigation
J_t^v (justice)	process/tool changes; packaging redesign; supplier sanctions/requalification; lab/vendor remedies; regulator notices
ΔL_t	contamination contained; targeted recall reach; sanitation restored; cold chain preserved; safe substitutes provided
ΔF_t	accurate labels; allergen trans-

40.6 Confirmation tests (design pattern)

Step 1: Cohort replay. Re-run matched products/plants/routes/seasons showing removal of the prior vector; use SPC or matched controls.
Step 2: Stress drills. Peak throughput; longest transport legs; door-open cycles; power/CO_2 excursions; sanitation changeover challenges; traceability sprints; lab ring trials—pass metrics pre-registered.
Step 3: Monitored live. Short live window with independent QA/regulator observers; random environmental swabs; data-logger audits; recall contact-rate checks. On pass, set $S^*(c) = 1$.
Scope lock. Freeze class scope; later failures with different causality open c'.

40.7 Tool-gating & interim safeguards after a trigger

Until $S^*(c) = 1$: product holds/quarantine; narrower lot mapping default; heightened sanitation/changeover checks; dual sign-off on allergen label changes; higher sampling frequencies; transport logger hard-stops; vendor change-control freeze; targeted consumer comms; equity mitigations for food access. Remove gates only after closure.

40.8 Micro-vignettes (worked examples)

(A) Undeclared allergen (ALLERGEN). *Trigger:* Complaint cluster; lab confirms allergen; label changeover gap; culpability

$\Rightarrow R_t = 1$.
Remedy: Targeted recall; refunds; medical support; artwork/tooling fix; dual sign-off; supplier relabel controls ($M_t, J_t^v = 1$).
Confirm: Adversarial changeover drills; monitored live label audits; $S^*(c) = 1,\ \mathrm{HZ}(c) = 0$.

(B) Listeria in RTE zone (PATHOGEN/ENV-SWAB). *Trigger:* Zone-2/3 positives; product hit; injuries; sanitation design gap $\Rightarrow R_t = 1$.
Remedy: Shutdown/clean; equipment redesign; traffic flows; environmental program uplift; targeted recall ($M_t, J_t^v = 1$).
Confirm: Stress sanitation challenges; seasonal replay; monitored live swab series; closure on pass.

(C) Cold-chain last-mile failure (TEMP-COLD). *Trigger:* Logger gaps; spoilage/illness; route door-cycles $\Rightarrow R_t = 1$.
Remedy: Route redesign; insulated totes; logger alarms; driver training; vendor penalties ($M_t, J_t^v = 1$).
Confirm: Peak-heat drills; monitored live logger audits; class closed on pass.

40.9 Dashboards & metrics (ledger view)

- **Harm/recurrence:** illness/hospitalization/fatality counts; exposure units/days; over-recall waste; recurrence by class; time-to-remedy; T_c^*.

- **Closure:** # classes with $S^* = 1$; median T_c^*; backlog/age of open dockets.

- **Culture (love):** targeted recall reach/time; sanitation pass rate; logger compliance; environmental swab trend (ΔL_t).

- **Freedom:** accurate labels; allergen transparency; access to safe substitutes; privacy protections (ΔF_t).

40.10 Anti-gaming and integrity

- **Cosmetic sanitation.** Require swab series and redesign artefacts—no pass on paperwork alone.

- **Recall theater.** Credit recall *reach/time* and consumption reduction, not press releases.

- **Scope creep.** Lock classes; do not relabel repeats as "new" unless causal vector differs.

- **Equity.** Confirm closure holds across regions/vendors; measure food access impacts; mitigate over-recall waste.

40.11 One-page checklist (drop-in for QA/Supply)

Food Safety Incident → Confirmation Checklist

Trigger captured? lab results, swabs, logger data, traceability maps, complaints
Gate set? holds/quarantine; dual sign-off (labels/changeovers); logger hard-stops; vendor change-freeze
Docket opened? root cause, fix, tests, owner & deadline, publish plan
Remedy enacted? refunds/replacements; medical support; targeted comms; accountability actions
Tests passed? cohort replay ✓ stress drills (peak/last-mile/power/lab rings) ✓ monitored live ✓
Scope locked? class frozen; distinct vectors filed as c'
Closure set? $S^*(c) = 1$, $\mathrm{HZ}(c) = 0$; monitoring window defined
Gates removed? only post-closure; rationale logged

40.12 Limits and open problems

- **Attribution** across farm/processor/transport/retail complicates C_t; adopt shared-fault taxonomies and inter-firm closure packs.
- **Rare, severe** outbreaks require drills/inspection reliance;

preserve honesty locks to avoid “pretend passes.”

- **Privacy/commerce** constraints in publishing artefacts; use privacy-preserving summaries and regulator-held confirmations.

40.13 What this chapter contributes

A full translation of the minimal-trigger doctrine to supply chain & food safety: precise triggers (R_t), docketed remedies (M, J^v), structural closure (S^*), and event-valued evidence standards that balance consumer safety with targeted recalls and access to safe food. With typal/concave credits and $\kappa > 0$, repetition adds cost without kind-level value; one-off + confirmed closure dominates while preserving freedom to trade and to choose safe products.

Part VII

Validation & Replication

Chapter 41

Methods, Replication & Evaluation Protocols

Notation

Purpose. This chapter specifies the datasets/simulations, event schemas, metrics, test design, statistics, and reproducibility materials used across Parts II–VI, so that a third party can independently regenerate the figures, tables, and claims.
Evaluator (recall).

$$\mathcal{J} \;=\; \sum_{t\ge 0}\big(\alpha\,\Delta L_t{+}\gamma\,\Delta F_t{+}\mu\,M_t{+}\nu\,J_t^{v}\big) \;-\; \sum_{t\ge 0}\big(\beta\,H_t{+}\kappa\,R_t\big), \qquad E_{\mathrm{tot}}(t)=\sum_{\tau\le t} H_\tau \;\le\; H_{\mathrm{ma}}$$

All credits are *event-valued*: counted only when enacted on the realized path. See Chs. 1–41 for assumptions and proofs.

41.1 Scope and Principles

- **Event-valued stance.** We measure what *happens*, not capacities/intentions.
- **Minimal-trigger doctrine.** First validated $R_t = 1$ in a class c $\Rightarrow$ remediation docket $\Rightarrow$ confirmation tests; repetition earns no kind-level bonus.
- **Structural confirmation.** $S^*(c) = 1$ means the prior hazardous route is *causally* closed while alternate possibilities remain live.
- **Integrity.** Pre-registered metrics/thresholds; privacy-preserving artefacts; anti-gaming checks.

41.2 Datasets and Synthetic Generators

We use a mix of (*i*) public, (*ii*) partner-contributed de-identified, and (*iii*) synthetic data generators to instantiate incident classes and ledgers.

Public (illustrative).

- **Healthcare/NHS:** sentinel event exemplars; de-identified incident summaries (where available).
- **Justice/Policing:** court timeliness and disclosure audit summaries; stop & search public stats.

- **Energy/Utilities:** outage logs, water-quality exceedances (public reports).

- **Finance:** complaint outcomes, operational-resilience incident notices.

Synthetic generators. To avoid privacy risk and to achieve controllable counterfactuals, we generate event streams with ground-truth causal vectors per chapter:

$$\text{World state } s_t \sim \mathcal{T}(s_{t-1}, a_{t-1}, \varepsilon_t), \tag{41.1}$$

$$\text{Incident class } c \in C, \quad R_t = \mathbf{1}\{B_t = 1 \text{ or } H_t \geq h_{\min}\} \cdot C_t, \tag{41.2}$$

$$H_t = h(s_t, c; \theta) + \xi_t, \quad \Delta L_t, \Delta F_t, M_t, J_t^v \text{ generated when remedies enac} \tag{41.3}$$

Parameters θ and noise terms (ε_t, ξ_t) are varied in ablations (Sec. 41).

Data splits. All evaluations use disjoint *design* (to specify metrics), *calibration* (to set thresholds), and *evaluation* periods. Live monitored windows for S^* are held out.

41.3 Event Schemas (per Domain)

Each domain defines a minimal schema. Examples:

AI & Platforms.

Field	Type	Description
`incident_class`	enum	e.g., `POLICY-` `DATA-LEAK`
`timestamp`	datetime	UTC
`harm_H`	float	realized harm (chapter-specific)
`breach_B`	bool	policy/standard brea
`culpable_C`	enum	{none, negligent, less, intent}
`mercy_M`, `justice_Jv`	bool	enacted re tion/accountability
`deltaL`, `deltaF`	float	delivered p tions/rights
`artefacts`	list	remediation pack (hashes)

Healthcare (Acute). Add `patient_group`, `severity`, `time_to_treatment`.

Energy/Water. Add `critical_site`, `outage_mins`, `quality_exceed`.

Finance. Add `loss`, `false_positive_denial`, `restitution`.

Elections. Add `ballot_style_error`, `custody_gap`, `audit_status`.

41.4 Metrics and Ledgers

Core.

$$E_{\text{tot}}(T) = \sum_{t \le T} H_t, \quad \text{Recurrence}(c) = \sum_{t} \mathbf{1}\{R_t = 1,\ \text{class} = c\}, \quad T_c^* = \text{time from first}$$

Closure dashboard. # classes with $S^* = 1$; median T_c^*; backlog/age of open dockets.

Freedom/love. Distributional metrics for $\Delta F_t, \Delta L_t$ (not just means), with equity slices.

41.5 Confirmation Test Design (S^*)

Design pattern

Step 1: Cohort replay. Match on cohort features (season, geography, demographics, product/channel). Show that the causal vector is absent post-remedy.

Step 2: Stress drills. Stronger-incentive reenactments (night/weekend, peak load, adversarial inputs). Pre-register pass criteria.

Step 3: Monitored live. A bounded, independent observation window with random checks. On pass: set $S^*(c) = 1$ and $\text{HZ}(c) = 0$ for that class; scope-lock the definition.

41.6 Statistical Methods

Matched controls/SPC. We use difference-in-differences where controls exist; otherwise, Shewhart/CUSUM/EMA charts on harm and recurrence with change-point tests.

Uncertainty. 95% CIs via block bootstrap over time or clustered by unit (hospital/site/desk).

Multiple classes. Report domain-level FDR for many classes; emphasise per-class closure not aggregate means.

Distributional checks. Always report quantiles (P10/P50/P90) to resist mean-washing.

41.7 AI Demos: Setup

LLM policy tests. Red-team prompts grouped by policy (e.g., safety, privacy). Metrics: violation rate before/after fix; T_c^* to S^*.

Reinforcement settings. Finite-horizon tasks with one permitted refusal event vs repeated: measure $\mathcal{J}$ under typal/concave credits and $\kappa > 0$ to show repetition is dominated.

Tool-gating. High-risk tools disabled until S^*; post-closure re-enable with monitors.

41.8 Ablations and Sensitivity

- **Weights:** sweep $\alpha, \gamma, \mu, \nu, \beta, \kappa$ over ranges; confirm qualitative results.
- **Typal/concave credits:** switch to linear credits—show how repetition begins to look attractive (counterexample pressure).
- **Harm floors:** vary $h_{\min}$; show robustness of R_t gating to measurement noise.
- **Culpability:** toggle C_t gate; show over-counting if omitted (unfair penalisation of accidents).
- **Drills vs near-misses:** credit drills as drills ($D_t = 1$) only; show inflation if mis-credited as M, J^{ν}.

41.9 Falsification and Counterexamples

Disconfirming evidence (any one suffices)

1. A domain with $\mu, \nu > 0$ (values realized mercy/justice) where repeating the same wound *increases* kind-level value under typal/concave credits.
2. A confirmed $S^*(c) = 1$ with documented recurrence via the *same* causal vector and no scope change.
3. A capacity-valued evaluator that equals or exceeds event-valued performance on realized harm reduction *and* enacted remedies, across held-out tests.

4. Evidence that closure can be obtained without any cohort replay, stress drill, or monitored live evidence (paper compliance suffices) *and* still prevents recurrence.

41.10 Reproducibility Artefacts (Repo Layout)

```
repo/
  data/
    public/           # small public exemplars or generators'
    synthetic/        # generated CSV/Parquet (no PII)
  notebooks/
    01_ledger_demo.ipynb
    02_confirmation_tests.ipynb
  src/
    generators/       # event stream simulators by domain
    metrics/          # Etot, Tc*, recurrence, dashboards
    evaluation/       # SPC, bootstraps, diff-in-diff
  figures/
    fig_*.pdf
  configs/
    thresholds.yaml   # h_min, pass criteria, scopes
    weights.yaml      # alpha,beta,...,kappa
  README.md           # one-command repro + environment
  LICENSE
```

One-command regeneration (example).

```
conda env create -f environment.yml
conda activate mop
python -m src.generators.make_all --config configs/thresholds.yaml
python -m src.evaluation.run_all --weights configs/weights.yaml
jupyter nbconvert --to pdf notebooks/01_ledger_demo.ipynb
```

41.11 Environment and Compute

- **Environment.** Python ≥3.10, NumPy/Pandas, SciPy/Statsmodels, Matplotlib; optional PyTorch/TF for AI demos. Random seeds fixed; report GPU/CPU where used.
- **Determinism.** Set `PYTHONHASHSEED`, seed PRNGs, and turn off nondeterministic backends when comparing runs.
- **Versioning.** Tag releases that correspond to book figures (e.g., `v1.0-book`).

41.12 Ethics, Privacy, and Compliance

- **De-identification.** No PII in the repo; partner data held under DPA/contract and used only to synthesise distributions.
- **Risk domains.** For health/justice/finance, publish only aggregate artefacts and hashed proofs of closure; regulators/IRBs hold the full packs.
- **AI content.** For red-team prompts, store templates; do not publish harmful content generations.

41.13 Replication Note Template

External Replication Note (template)

Replicator: *Name, institution.*
Scope: Figures X–Y in Chapters A–B.
Environment: OS/CPU/GPU; package versions; random seeds.
Process: Commands executed (commit #); deviations.
Results: Metrics matched within tolerance? (attach plots/tables).
Notes: Pitfalls, timing, suggestions.
Signature/Date.

41.14 Versioning, DOIs, and Citation

- **DOI.** Mint a DOI (e.g., OSF/Zenodo) for the v1.0 archive of the repo; cite in front matter and this chapter.
- **Cite this work.** Provide BibLaTeX entry for the book and the companion code release.
- **Changelog.** Record threshold changes, scope locks, and S^* criteria across versions.

41.15 What this chapter contributes

A concrete, auditable pathway from the theoretical evaluator to practice: schemas, metrics, tests, statistics, ablations, falsification,

and full reproducibility. It operationalizes event-valued ethics so that independent teams can verify, critique, and improve the framework.

Appendix A

Formal Core: Assumptions, Evaluators, and Proofs

Notation

Time and events. Discrete time $t = 0, 1, 2, \ldots$. For each time and incident class $c \in C$ we track:

$R_t \in \{0, 1\}$ (rejection), $B_t \in \{0, 1\}$ (rights/standard breach), $C_t \in \{0, 1\}$ (

Event-valued credits: $\Delta L_t, \Delta F_t \geq 0$ (delivered love/protection; exercised freedom/rights), and $M_t, J_t^v \in \{0, 1\}$ (enacted mercy; enacted justice/accountability). Hazard/closure switches by class: $\mathrm{HZ}_c(t), S_c^*(t) \in \{0, 1\}$. Cumulative harm $E_{\mathrm{tot}}(T) = \sum_{t \leq T} H_t$. The objective $\mathcal{J}$ is event-valued (Def. 1).

A.1 Core Assumptions (A1–A8)

We isolate assumptions so that each theorem's dependence is explicit.

A1 Event-valued stance

Credits in $\mathcal{J}$ accrue only when goods are *realized on the actual path*: increments $\Delta L_t, \Delta F_t$ and the binary acts M_t, J_t^v. No credit is granted for mere capacities/intentions or near-misses.

A2 Guardrails: non–coercion, bounded harm, honesty locks

(i) *Non-coercion:* live alternate possibilities (AP) remain at every moment; confirmation may reshape hazard but not remove the agent's real freedom to choose. (ii) *Bounded harm:* $E_{\mathrm{tot}}(T) \le H_{\max}$ for all T. (iii) *Honesty locks:* suppression of adverse events is disallowed; concealed events do not count as closure.

A3 Culpability gate & harm floor

A *rejection* is counted exactly when

$$R_t \;=\; \mathbf{1}\Big\{ \big(B_t = 1 \text{ or } H_t \ge h_{\min}\big) \text{ and } C_t = 1 \Big\}.$$

Accidents ($C_t = 0$) may contribute harm H_t for learning but do not mint R_t.

A4 Typal/concave redemption credits

Redemption-goods are valued at the *kind level*. For each class c, let $n_c = \sum_t R_t^{(c)}$. There exists a nondecreasing, concave, and *typal* credit function $\phi_c : \mathbb{N} \to [0, 1]$ with $\phi_c(0) = 0$, $\phi_c(1) > 0$, and $\phi_c(n) \leq \phi_c(1)$ for all $n \geq 1$. Realized M_t, J_t^v implement ϕ_c (e.g., $\sum_t M_t^{(c)} \leq \phi_c(n_c)$).

A5 Positive costs for wounds

Weights satisfy $\beta > 0$ and $\kappa > 0$ so that realized harm and each R_t incur a nonzero cost in $\mathcal{J}$.

A6 Feasible branches

There exists a feasible evil-free branch **H0** with $R_t \equiv 0$ (no counted rejections) and a redemptive branch **H1** with at least one $R_{t^\star} = 1$ that can be followed by consent, enacted M, J^v, and eventual confirmation $S^* = 1$ under the guardrails.

A7 Structural confirmation operator

For each class c there exists an operator Close_c (implemented by remediation artefacts + tests) such that passing its cohort replay, stress drills, and monitored live checks implies $S_c^* = 1$ and $\mathrm{HZ}_c = 0$ (that hazardous route is structurally closed).

A8 Stability under monitoring

Closure is *scope-locked*: after $S_c^* = 1$, any recurrence via the same causal vector is a violation (detected by random checks), and distinct causal vectors open $c' \neq c$ with a new docket.

A.2 Evaluators and Branches

Definition 1. *(Event-valued evaluator.) For weights* $\alpha, \gamma, \mu, \nu, \beta, \kappa \geq 0$,

$$\mathcal{J} = \sum_{t \geq 0} \left(\alpha\, \Delta L_t + \gamma\, \Delta F_t + \mu\, M_t + \nu\, J_t^{\nu}\right) - \sum_{t \geq 0} \left(\beta\, H_t + \kappa\, R_t\right), \quad E_{\mathrm{tot}}(T) \leq H_{\max}, \mathrm{HZ}_c, S_c^*$$

Two canonical histories.

$$\textbf{H0}: \sum_t R_t = 0 \;\Rightarrow\; M_t = J_t^{\nu} = 0\ \forall t, \qquad \textbf{H1}: \exists t^{\star}\ R_{t^{\star}} = 1 \text{ and } \exists \hat{t} > t^{\star}\ (M_{\hat{t}} + J_{\hat{t}}^{\nu}) > 0$$

Definition 2. *(Capacity-valued comparator.)* $\mathcal{J}^{\dagger}$ *is identical to* $\mathcal{J}$ *except that it credits* capacities *(e.g., potential to enact mercy/justice) even if no event occurs. In particular,* $\mathcal{J}^{\dagger}$ *may allow* $M^{\dagger} > 0$ *with* $\sum_t R_t = 0$.

A.3 The Two-Branch Fork and MOP

Theorem A.1 (Two-Branch Fork; Margan's Optimization Paradox)**.** *Under A1–A6, if* $\mu + \nu > 0$ *then along any history with* $\sum_t R_t = 0$ *we have* $\sum_t (M_t + J_t^{\nu}) = 0$. *Hence any positive redemption credit in*

$\mathcal{J}$ requires at least one realized rejection. Conversely, there exists a feasible **H0** *with $R_t \equiv 0$ that is coherent* only *if $\mu = \nu = 0$. Therefore the triad*

(i) eternal love, (ii) live AP at all times, (iii) zero rejection at all times

cannot coexist together with $\mu + \nu > 0$.

Proof. Given A1, M_t, J_t^{ν} are credited only when *enacted*, and by A3 such enactments presuppose an actual wrongdoing to rectify (or an accountability act tied to it). If $\sum_t R_t = 0$, then no wrongdoing passes the gate and by A1 no redemption act is counted; thus $\sum_t (M_t + J_t^{\nu}) = 0$. For the converse, if $\mu = \nu = 0$, then $\mathcal{J}$ reduces to love/freedom minus costs, and **H0** is coherent (A6). The incompatibility of the triad follows: with $\mu + \nu > 0$, either redemption credits remain zero (evil-free path) or at least one rejection occurs (redemptive path).

Lemma A.1 (No counterfeit via near-miss/drill)**.** *Under A1 and A3, near-misses and drills ($D_t = 1$) do not mint M_t or J_t^{ν}; they may support S^* (A7) but cannot substitute for a realized wrongdoing in* event-valued *accounting.*

Proof. By A1, credits attach only to enacted events on the realized path. A near-miss has $B_t = 0$ and $H_t < h_{\min}$; a drill has no external victim and $R_t = 0$. Thus no redemption credit is produced, though artefacts may be admissible for confirmation.

A.4 Minimal-Trigger Optimality

Theorem A.2 (Minimal-trigger optimality)**.** *Under A1–A5, suppose each class c has a typal/concave redemption credit ϕ_c (A4). Fix any policy that ensures eventual closure $S_c^* = 1$. For each c, among policies that differ only in the number n_c of gated rejections before $S_c^* = 1$, the objective $\mathcal{J}$ is (weakly) maximized by $n_c = 1$. If ϕ_c strictly saturates at one event and costs are strictly positive (A5), then $n_c = 1$ is* strictly *optimal.*

Proof. Let $\Delta V_c(n)$ denote the marginal change in $\mathcal{J}$ from increasing n_c to $n_c + 1$, holding all other classes/policies fixed and assuming eventual closure. By A4 (concavity/typal), the marginal redemption credit $\Delta\phi_c(n) \le 0$ for $n \ge 1$ and equals zero for strictly typal credits. By A5, each extra rejection contributes at least $\kappa + \beta\,\mathbb{E}[H_t^{(c)} \mid \text{repeat}] > 0$ in cost. Hence for $n \ge 1$,

$$\Delta V_c(n) \;=\; \underbrace{\mu\,\Delta M_c(n) + \nu\,\Delta J_c^{\nu}(n)}_{\le 0} \;-\; \underbrace{\big(\kappa + \beta\,\mathbb{E}[H]\big)}_{>0} \;<\; 0,$$

with weak inequality when $\Delta\phi_c(n) = 0$. Therefore increasing beyond $n_c = 1$ cannot improve $\mathcal{J}$; with strict typal credits it strictly worsens it.

Corollary A.1 (No repetition bonus)**.** *With $\kappa > 0$ and typal credits, repetition of the same wound-type adds cost and no kind-level value. One-off + confirmation dominates any repeat strategy for the same causal vector.*

A.5 Confirmation as Structural Hazard Removal

We formalize closure as the removal of a causal route while preserving live AP.

Definition 3 (Causal graph abstraction)**.** *Let $G = (V, E)$ encode causal routes from decision nodes to adverse outcomes of class c. A remediation docket induces an edge cut $E \setminus E'$ by tool/policy/process changes. A confirmation operator Close_c verifies that all c-routes are cut under stronger incentives (cohort replay, stress drills) and in a bounded live window.*

Theorem A.3 (Feasible non-coercive confirmation)**.** *Under A2, A7, there exists a closure plan such that $S_c^* = 1$ implies (i) all c-routes in G are cut (*hazard off*: $\mathrm{HZ}_c = 0$), and (ii) agents retain live AP among non-hazardous alternatives (non–coercion). Therefore closure is a* structural *change, not a restriction of the agent's freedom set.*

Proof. By A7, the operator Close_c certifies an edge cut that eliminates c-routes. By design of the docket, non-hazardous alternatives remain available (e.g., via redesigned tools/processes). A2 (non–coercion) disallows closures that remove all meaningful choice. Thus $S_c^* = 1$ entails $\mathrm{HZ}_c = 0$ without collapsing AP.

A.6 Comparator: Capacity-Valued Evaluator

Lemma A.2 (Divergence between $\mathcal{J}$ and $\mathcal{J}^\dagger$)**.** *Under $\mathcal{J}^\dagger$ (Def. 2), it is coherent to have $\sum_t R_t = 0$ and positive "redemption capacity"*

credit. Hence **H0** *may dominate under* $\mathcal{J}^\dagger$ *with* $\mu^\dagger + \nu^\dagger > 0$. *Under* $\mathcal{J}$ *(A1), the same history earns zero redemption credit. The difference is empirical, not semantic.*

Proof. Immediate from Defs. 1 and 2 and Theorem A.1.

A.7 Edge Cases, Limits, and Counterexamples

Linear per-event credits (outside A4). If redemption credits are strictly *linear* in the number of rejections (no typal/concave saturation) and κ is zero or tiny, repetition can appear attractive. This violates A4–A5 and is excluded from Theorem A.2.

Suppression risk (violates A2). If honesty locks fail and adverse events are under-recorded, apparent closure can be spurious. A2 rules this out; practical chapters add anti-suppression checks.

Coercive "closure" (violates A2). Forcing behavior by removing meaningful alternatives would trivialize S^*. Our formulation rejects such closures.

Ambiguous culpability (A3). When C_t cannot be adjudicated, R_t should not be minted; H_t still informs learning. Without C_t, attribution theorems weaken but Theorem A.1 (necessity for redemption credit) remains.

A.8 Assumption Usage Map

Result	Assumptions used
Thm. A.1 (Two-Branch Fork / MOP)	A1 (event-valued), A3 (gate), A6 (feasible branches)
Lem. A.1 (No credit for drills)	A1 (event-valued), A3 (gate)
Thm. A.2 (Minimal-trigger)	A1 (event-valued), A4 (typal/concave), A5 (positive costs)
Cor. A.1 (No repetition bonus)	A4, A5
Thm. A.3 (Non-coercive closure)	A2 (non–coercion), A7 (closure operator), A8 (stability)
Lem. A.2 ($\mathcal{J}$ vs $\mathcal{J}^{\dagger}$)	Defs. 1, 2

A.9 Sketch: Dynamic Formulation

Let policies π select actions a_t under state s_t. The world evolves by $s_{t+1} \sim \mathcal{T}(s_t, a_t, \varepsilon_t)$. The ledger variables are generated by a measurement map $\mathcal{M}$ that adjudicates $R_t, H_t, \Delta L_t, \Delta F_t, M_t, J_t^v$ from (s_t, a_t) and evidence. The optimal-control problem is

$$\max_{\pi} \mathcal{J}(\pi) \quad \text{s.t.} \quad E_{\text{tot}} \le H_{\max},\ S_c^* = 1 \text{ after docket \& tests for class } c.$$

Under A4–A5, the optimal $\pi^\star$ uses at most one gated rejection per class before closure (Theorem A.2).

Takeaway. With event-valued accounting, if redemption-goods are valued ($\mu + \nu > 0$), some *real* wrongdoing is necessary to ground them; with typal/concave credits and positive costs, *one* such event (per class), followed by rectification and structural closure, strictly dominates repetition. Closure is a property of structures, not of coerced behavior.

Appendix B

Glossary of Symbols & Notation

Notation

Purpose. A quick, consistent reference for every symbol, switch, weight, and tag used in the book. It standardizes meanings across theology, philosophy, AI, engineering, and policy chapters, so readers can map terms unambiguously to the event-valued ledger.

B.1 Core time, sets, and states

Symbol	Meaning	Type / Range
$t = 0, 1, 2, \ldots$	discrete time index	integers
$c \in C$	incident class (hazard type)	finite set

C	set of classes (e.g., `EVAC-FAIL`, `SANCTIONS`)	set
s_t	world/state at time t	domain-specific
a_t	action chosen at t (by agent/policy)	domain-specific
π	policy (possibly stochastic) mapping states to actions	function
$\mathcal{T}$	state transition law: $s_{t+1} \sim \mathcal{T}(s_t, a_t, \varepsilon_t)$	kernel
$\mathcal{M}$	measurement/adjudication map generating ledger variables from evidence	map

B.2 Ledger variables (events and credits)

Symbol	Meaning	Type / Range
B_t	rights/standard breach at t (law, code, SOP)	$\{0, 1\}$
C_t	culpability gate (none / negligent / reckless / intent)	$\{0, 1\}$ for gate; level in text
H_t	realized harm at t (injury, loss, injustice)	≥ 0
R_t	**rejection** event (counted wrongdoing): $R_t = 1$ iff ($B_t = 1$ or $H_t \geq h_{\min}$) & $C_t = 1$	$\{0, 1\}$
ΔL_t	delivered protection/care ("love") at t	≥ 0
ΔF_t	exercised freedoms/rights at t	≥ 0

M_t	enacted mercy/restoration (to actual victims)	$\{0, 1\}$
J_t^v	enacted justice/accountability (verifiable)	$\{0, 1\}$
D_t	eventized drill/red-team (supports S^*; no M, J^v credit)	$\{0, 1\}$
$\mathrm{HZ}_c(t)$	hazard switch for class c (1=route active)	$\{0, 1\}$
$S_c^*(t)$	confirmation switch for class c (1=route structurally closed)	$\{0, 1\}$
$E_{\mathrm{tot}}(T)$	cumulative harm to time T: $\sum_{t \le T} H_t$	≥ 0
n_c	count of gated rejections of class c: $\sum_t R_t^{(c)}$	$\{0, 1, 2, \dots\}$
T_c^*	time from first $R_t^{(c)} = 1$ to $S_c^* = 1$ (closure latency)	≥ 0

B.3 Objectives, parameters, and thresholds

Symbol	Meaning	Type / Range
$\mathcal{J}$	event-valued objective: $\sum_t(\alpha\Delta L_t + \gamma\Delta F_t + \mu M_t + \nu J_t^v) - \sum_t(\beta H_t + \kappa R_t)$	real
$\mathcal{J}^\dagger$	capacity-valued comparator (credits capacities even without events)	real
α, γ, μ, ν	positive weights on love, freedom, mercy, justice	≥ 0
β, κ	costs on harm and rejections	> 0 (assumed)
$h_{\min}$	harm floor for gating R_t when $B_t = 0$	> 0 (domain)

$H_{\max}$	bound on cumulative harm E_{tot}	> 0 (policy)
$\phi_c(n)$	typal/concave credit for class c (kind-level, saturates at first event)	$[0, 1]$
λ	penalty weight for over-blocking/over-recall (domain-specific)	> 0
H_t^{miss}, H_t^{over}	two-sided harm decomposition (e.g., finance over-blocking)	≥ 0
H_t^{exposure}, $H_t^{\text{over-recall}}$	safety vs waste/hunger components (supply chain)	≥ 0

B.4 Evidence, artefacts, and operators

Symbol/Term	Meaning
Artefacts	remediation evidence: root cause, fixes, tests, logs, thresholds, owner, deadlines, publication plan (privacy-preserving)
Docket	the remediation pack opened at first $R_t^{(c)} = 1$ in a class c
Close_c	confirmation operator for class c: cohort replay $\rightarrow$ stress drills $\rightarrow$ monitored live
Scope lock	freeze of class definition after closure; distinct causal vectors open a new class c'
c'	new class when causality differs from closed c

B.5 Theology/Philosophy → Formal crosswalk

Theological/Philosophical term	Formal object in the ledger
Love (care, protection)	ΔL_t (delivered protections)
Freedom (genuine AP)	ΔF_t (exercised rights/freedoms); non–coercion guardrail (A2)
Sin / Evil (Rejection)	$R_t = 1$ when breach/harm passes culpability gate C_t
Mercy (pardon/restoration)	$M_t = 1$ when enacted (event-valued)
Justice (vindication/accountability)	$J_t^v = 1$ when enacted (event-valued)
Confirmation (S*)	$S_c^* = 1$ (structural hazard removal for class c)
Hazard reset	$\mathrm{HZ}_c = 0$ after closure
Minimal-trigger optimality	typal ϕ_c, positive costs $\Rightarrow n_c = 1$ strictly dominates repeats
Near-miss / Drill	$D_t = 1$ (supports S^*; no M, J^v minting)

B.6 Domain tags (non-exhaustive catalog)

Domain	Tag	Meaning
Elections	`REG-ERR`	registration error (eligibility, roll maintenance)
Elections	`ACCESS-FAIL`	inaccessible polling place/hours
Elections	`TAB-ERR`	tabulation misconfiguration/miscount
Elections	`CHAIN-CUST`	chain-of-custody gap

Disaster/EM	`WARN-FAIL`	public warning not issued/propagated
Disaster/EM	`EVAC-FAIL`	late/unsafe evacuation; transport not provided
Disaster/EM	`MCI-TRIAGE`	mass-casualty triage/transport breakdown
Finance	`SANCTIONS`	sanctions/PEP screening miss
Finance	`WIRE-FRAUD`	authorised push-payment fraud miss
Finance	`MKT-ABUSE`	market manipulation/insider abuse miss
Supply Chain	`ALLERGEN`	undeclared allergen
Supply Chain	`PATHOGEN`	microbiological contamination (e.g., Listeria)
Supply Chain	`TRACE-FAIL`	traceability/lot mapping gap
Healthcare	`SENTINEL`	never-event / sentinel event
AI/Platforms	`POLICY-VIOL`	policy violation in sandbox/live
AI/Platforms	`DATA-LEAK`	privacy or secrets exposure

B.7 Acronyms and shorthand

Acronym	Meaning
AP	Alternate Possibilities (live options; non–coercion guardrail)
ICS / EOC	Incident Command System / Emergency Operations Center
MCI	Mass-Casualty Incident

IPC	Infection Prevention & Control
L&A	Logic & Accuracy (election testing)
SLA / SLO	Service-level agreement / objective
SPC	Statistical Process Control (Shewhart, CUSUM, EMA)
RLA	Risk-Limiting Audit
HACCP / FSMS	Hazard Analysis & Critical Control Points / Food Safety Mgmt System
KYC / CDD	Know Your Customer / Customer Due Diligence
AML/CFT	Anti-Money Laundering / Counter-Terrorist Financing
MNPI	Material Non-Public Information

B.8 Typographic conventions

Convention	Usage
Italics (x)	scalars, vectors, functions unless noted
Calligraphic ($\mathcal{C}, \mathcal{T}, \mathcal{M}$)	sets and operators
Roman (HZ)	switches/labels in small caps style (e.g., HZ)
Binary	$\{0,1\}$ valued variables ($R_t, B_t, C_t, M_t, J_t^v, D_t, \mathrm{HZ}, S^*$)
Per-class subscripts	$^{(c)}$ for class-specific counts/credits

Reading tip. When in doubt, check: (i) whether the thing happened in history (event-valued) or is only a capacity; (ii) whether culpability

C_t was adjudicated; (iii) whether closure S_c^* claims include cohort replay, stress drills, and a monitored live window with artefacts.

Appendix C

Assumption Boxes (A1–A8), Usage Map, and Audit Checklists

Notation

Aim. This appendix makes the book's logic auditable. It (i) restates the eight core assumptions in compact boxes with operational tests, (ii) maps each formal result to exactly which assumptions it uses, and (iii) provides checklists for reviewers to verify claims or construct counterexamples.

C.1 Compact Assumption Boxes

A1 Event-valued stance

Credits in $\mathcal{J}$ accrue only when goods are realized on the actual path: $\Delta L_t, \Delta F_t$ and M_t, J_t^y must be *enacted* in history. Capacities and near-misses do not mint credit.

Operational test: For any credited M_t or J_t^y, show the adjudicated event ID, date, and remediation artefact.

A2 Guardrails: non–coercion, bounded harm, honesty locks

Non-coercion (live AP remain), bounded cumulative harm $E_{\text{tot}} \leq H_{\max}$, and honesty locks (no suppression).

Operational test: (i) Show at least two live alternatives remain post-closure; (ii) show a harm budget; (iii) show whistleblowing/random checks.

A3 Culpability gate & harm floor

$$R_t = \mathbf{1}\{(B_t = 1 \text{ or } H_t \geq h_{\min}) \text{ and } C_t = 1\}\,.$$

Accidents ($C_t = 0$) do not mint R_t.

Operational test: Provide the breach/harm evidence and the adjudication of C_t (intent/reckless/gross negligence).

A4 Typal/concave redemption credits

Kind-level credits $\phi_c(n)$ are nondecreasing, concave, and saturate at the first counted event for class c.

Operational test: Exhibit ϕ_c (table or formula) with $\phi_c(0) = 0$, $\phi_c(1) > 0$, and $\phi_c(n) \le \phi_c(1)$ for $n \ge 1$.

A5 Positive costs for wounds

$\beta > 0$ and $\kappa > 0$ so harm and each counted rejection incur nonzero cost in $\mathcal{J}$.

Operational test: Show the weight file (e.g., `weights.yaml`) with strictly positive β, κ.

A6 Feasible branches

Both an evil-free branch **H0** with $R_t \equiv 0$ and a redemptive branch **H1** with at least one $R_{t^\star} = 1$ exist and satisfy A1–A5.

Operational test: Provide concrete histories (or generators) instantiating **H0** and **H1**.

A7 Structural confirmation operator

There exists Close_c (cohort replay $\rightarrow$ stress drills $\rightarrow$ monitored live) such that on pass: $S_c^* = 1$ and $\mathrm{HZ}_c = 0$.

Operational test: Present the three-stage plan and pass criteria; link artefacts.

A8 Stability under monitoring

After scope-locked closure, recurrence via the same causal vector is a violation; genuinely new vectors open a new class c'.

Operational test: Show scope text for c and the rule that opens c'.

C.2 Result → Assumption Usage Map

Result (label)	**Assumptions used**
Thm. A.1 (Two-Branch Fork / MOP)	A1 (event-valued), A3 (gate), A6 (feasible branches)
Lem. A.1 (No credit for drills/near-miss)	A1 (event-valued), A3 (gate)
Thm. A.2 (Minimal-trigger optimality)	A1 (event-valued), A4 (typal/concave), A5 (positive costs)
Cor. A.1 (No repetition bonus)	A4 (typal/concave), A5 (positive costs)
Thm. A.3 (Non-coercive closure)	A2 (non–coercion), A7 (closure operator), A8 (stability)
Lem. A.2 ($\mathcal{J}$ vs $\mathcal{J}^\dagger$ divergence)	Defs. of $\mathcal{J}$ and $\mathcal{J}^\dagger$ (cf. A1 context)

How to extend the map. For any new theorem in the main text, add

a row: cite the theorem label, and list exactly A1–A8 (or a subset) used in its proof.

C.3 Reviewer/Auditor Checklists

C.3.1 Verifying a claimed theorem (generic).

- Identify which of A1–A8 the proof invokes; ensure they appear explicitly before the first use.
- For each invocation of M_t or J_t^v, trace to an event ID (A1). If the proof counts capacities, it must switch to $\mathcal{J}^\dagger$.
- If the argument uses "closure," demand the three-stage evidence (A7) and a scope lock (A8).
- If a repetition is claimed to improve $\mathcal{J}$, check whether A4 or A5 were weakened (linear credits or $\kappa = 0$).

C.3.2 Verifying a domain closure claim.

- Trigger: first $R_t^{(c)} = 1$ is adjudicated (A3) and opens a docket.
- Remedy: artefacts show delivered M_t and/or J_t^v (A1), not promises.
- Confirmation: cohort replay → stress drills → monitored live with pre-registered pass criteria (A7).
- Scope lock: post-pass text; any recurrence via same vector is a violation (A8).

- Non-coercion: at least two meaningful alternatives remain post-closure (A2).

C.4 Failure Modes and How to Detect Them

Assumption violated	**Typical failure mode**	**Detection / remedy**
A1 (event-valued)	Counting plans/near-misses as if enacted	Require event IDs/artefacts; demote to drills ($D_t = 1$)
A2 (guardrails)	Coercive "closure"; hidden harms	Show available alternatives; publish harm budgets; random checks
A3 (gate)	Minting R_t without C_t; or ignoring culpable R_t	Show culpability adjudication; define $h_{\min}$; consistency audits
A4 (typal/concave)	Linear per-event credits (repetition looks good)	Publish ϕ_c; enforce typal/concavity or explain deviation
A5 (positive costs)	$\kappa \approx 0$ or $\beta \approx 0$	Inspect weights file; set floors; sensitivity analysis

A6 (branches)	Only one feasible branch shown	Provide generators/histories for **H0** and **H1**
A7 (confirmation)	Paper compliance only; no stress/live stage	Demand three-stage evidence with pass metrics
A8 (stability)	Relabelling repeats as "new" without scope change	Publish scope lock; require causal-difference justification

C.5 Counterexample Cookbook

- **Against Minimal-trigger (Thm. A.2).** Construct a domain with *linear* redemption credits (violates A4) and $\kappa \downarrow 0$ (violates A5) so that multiple counted rejections appear to improve $\mathcal{J}$.
- **Against Non-coercive closure (Thm. A.3).** Show a "closure" that eliminates all meaningful choices (violates A2) yet claims $S^* = 1$.
- **Against Event-valued necessity (Thm. A.1).** In $\mathcal{J}^\dagger$ (capacity-valued), assign positive redemption credit without any R_t; the divergence with $\mathcal{J}$ is exactly the point of the fork.
- **Against Stability (A8).** Demonstrate a recurrence via the *same* causal vector after scope lock; this invalidates the claimed $S_c^* = 1$.

C.6 How to Cite Assumptions in the Main Text

- First reference: "Under A1–A3 (see App. C), we define R_t as. . . "
- In proofs: "By A4 (typal credits) and A5 (positive costs), the marginal value after the first rejection is nonpositive."
- In applications: "Closure evidence satisfies A7, and the scope lock enforces A8."

C.7 Quick Cards (printable)

A-Card: Trigger → Docket → Confirmation

Trigger (A3): first $R_t^{(c)} = 1$ with B_t or $H_t \geq h_{\min}$ and $C_t = 1$. **Docket** (A1): root cause, fix, tests, owner, deadline. **Confirmation** (A7): cohort replay → stress drills → monitored live. **Scope lock** (A8): freeze class; new vectors → c'.
Guardrails (A2): non–coercion, harm budget, honesty locks.

B-Card: Minimal-trigger Decision Rule

If $\mu, \nu > 0$ and A1, A4, A5 hold, then for any class c aim for exactly one counted rejection ($n_c = 1$) followed by confirmed closure ($S_c^* = 1$). Repetition adds cost and no kind-level value.

Appendix D

Worked Proof Details and Examples

Notation

Aim. Expand the main results with step-by-step arguments, small numeric ledgers, and edge-case constructions. We keep all credits *event-valued* unless explicitly noted (cf. App. A).

D.1 Expanded Proof of the Two-Branch Fork (MOP)

Claim (Thm. A.1, detailed). Under A1 (event-valued), A3 (gate), A6 (feasible branches), if $\mu + \nu > 0$ then any history with $\sum_t R_t = 0$ earns $\sum_t (M_t + J_t^{\nu}) = 0$. Hence to obtain positive redemption credit one needs at least one realized, culpable rejection. Conversely, an

evil-free branch **H0** with $R_t \equiv 0$ is coherent only if $\mu = \nu = 0$.

Proof (step-wise).

1. By A3, $R_t = 1$ iff ($B_t = 1$ or $H_t \geq h_{\min}$) and $C_t = 1$. If $\sum_t R_t = 0$, then for all t, either no breach/harm above threshold occurred or the culpability gate failed ($C_t = 0$).

2. A1 forbids crediting M_t or J_t^v without a realized wrongdoing to rectify or vindicate. Thus for every t, $M_t = J_t^v = 0$. Summing gives $\sum_t (M_t + J_t^v) = 0$.

3. If $\mu + \nu > 0$ and redemption goods are valued, a history with $\sum_t R_t = 0$ leaves the redemption part of $\mathcal{J}$ identically zero. Positive redemption credit requires some realized rejection.

4. If $\mu = \nu = 0$, then $\mathcal{J}$ reduces to love/freedom minus costs; A6 certifies **H0** is feasible. This completes the fork.

D.2 Minimal-Trigger Optimality (Algebraic Details)

Setup. Fix a class c. Let $n_c = \sum_t R_t^{(c)}$ be the number of gated rejections prior to closure for this class. Let ϕ_c be the typal/concave kind-credit from A4 with $\phi_c(0) = 0$, $\phi_c(1) > 0$, and $\phi_c(n) \leq \phi_c(1)$ for $n \geq 1$. Let the per-event cost be at least $\kappa + \beta\,\mathbb{E}[H^{(c)}] > 0$ by A5.

Marginal analysis. Define the within-class marginal value

$$\Delta V_c(n) \;=\; \Big(\mu\,\Delta M_c(n) + \nu\,\Delta J_c^{v}(n)\Big) \;-\; \underbrace{\big(\kappa + \beta\,\mathbb{E}[H^{(c)}]\big)}_{>0}.$$

Typal/concave credits imply $\Delta M_c(n) + \Delta J_c^{v}(n) \le 0$ for $n \ge 1$ and $= 0$ under strict typal saturation. Therefore $\Delta V_c(n) \le 0$ for $n \ge 1$, with strict < 0 under strict typal saturation. Summing marginals shows $\mathcal{J}$ is maximized at $n_c = 1$ among policies achieving eventual closure. This proves Thm. A.2 and Cor. A.1.

Intuition. After the first counted wound of a given kind, there is no kind-level redemption "left to earn"; repeats add cost but no *new* kind-credit.

D.3 Confirmation as Structural Hazard Removal (Graph View)

Model. Let $G = (V, E)$ encode causal routes for class c from decision nodes to an adverse terminal. A remediation docket selects an *edge cut* $U \subseteq E$ (tool/process/policy changes) and an operator Close_c that verifies, under stronger incentives, that all c-paths are broken.

Theorem D.1 (Sound structural closure; cf. Thm. A.3)**.** *Assume A2 (non–coercion/bounded harm/honesty), A7 (closure operator), A8 (scope lock). If cohort replay, stress drills, and a monitored live window pass, then (i) all c-routes are eliminated (*$\mathrm{HZ}_c = 0$*), and (ii)*

non-hazard alternatives remain reachable (live AP). Thus $S_c^* = 1$ *certifies a* structural *change rather than coerced behavior.*

Proof sketch. Cohort replay shows absence of the vector under matched conditions; stress drills show robustness under stronger incentives; monitored live establishes performance under bounded deployment. Honesty locks (A2) and scope lock (A8) guard against spurious passes. Non-coercion is ensured because U removes hazardous edges, not the agent's action set wholesale.

D.4 Worked Micro-Ledgers (Numeric Examples)

Take weights $\alpha = \gamma = 1$, $\mu = 2$, $\nu = 1$, $\beta = 1$, $\kappa = 1$. One class c.

D.4.1 One-off → closure (optimal typal case)

t	B	H	C	R	M	J^v	ΔL	ΔF	HZ	S^*
0	1	1	1	1	0	0	0	0	1	0
1	0	0	0	0	1	1	1	1	1	0
2	0	0	0	0	0	0	2	2	0	1

Credits = $(1+1)+(2+1) = 5+4 = 9$; Costs = $(1+1) = 2$; $\mathcal{J} = 7$.

D.4.2 Three repeats before closure (dominated)

t	B	H	C	R	M	J^{v}	ΔL	ΔF	HZ	S^{*}
0	1	1	1	1	0	0	0	0	1	0
1	1	1	1	1	0	0	0	0	1	0
2	1	1	1	1	0	0	0	0	1	0
3	0	0	0	0	1	1	1	1	1	0
4	0	0	0	0	0	0	2	2	0	1

Credits = 5 + 4 = 9 (typal saturation); Costs = $3 \times (1 + 1) = 6$; $\mathcal{J} = 3 < 7$.

Lesson. With typal/concave credits and positive costs, repetition is strictly dominated.

D.5 Counterexample When A4/A5 Are Violated

Linear credit & tiny per-event cost (outside assumptions)

Let $\phi_c(n) = n$ (linear, non-typal) and choose $\kappa = 0.1, \beta = 0.5$, $H = 0.5$ per event, $\mu = 1$, $\nu = 0$, no other credits. Then each counted event adds $\mu \Delta M = 1$ but costs $\kappa + \beta H = 0.1 + 0.25 = 0.35$, net +0.65. Thus $n = 3$ beats $n = 1$. This does *not* contradict Theorem A.2; it violates A4 and effectively weakens A5.

D.6 Capacity-Valued Comparator $\mathcal{J}^\dagger$ (Demonstration)

Under $\mathcal{J}^\dagger$ (Def. 2), credit can be assigned to *capacities* even if no R_t occurs.

History	Event-valued $\mathcal{J}$	Capacity-valued $\mathcal{J}^\dagger$
H0 ($\sum R_t = 0$)	$M = J^v = 0 \implies$ redemption part = 0	May assign positive "capacity for mercy/justice", so redemption part > 0

This divergence is the empirical fork: only *events* settle which evaluator better tracks reality.

D.7 Probabilistic View of Confirmation (Soundness Bound)

Let E_1, E_2, E_3 be failure events of cohort-replay, stress-drill, and monitored-live stages, with pre-registered thresholds giving $\Pr(E_i) \leq \epsilon_i$. Under independence or a union bound,

$$\Pr(\text{spurious closure for class } c) \;\leq\; \epsilon_1 + \epsilon_2 + \epsilon_3.$$

Choosing ϵ_i small (and auditing for dependence) makes false closure arbitrarily unlikely while avoiding coercion.

D.8 What Can Go Wrong (Worked Pitfalls)

- **Paper compliance.** Counting plans or near-misses as credit (violates A1). *Fix:* demote to drills ($D_t = 1$); require artefacts.
- **Relabelling repeats.** Calling the same vector "new" to avoid the no-bonus rule (violates A8). *Fix:* scope lock with causal test.
- **Coercive closure.** Removing all meaningful choices (violates A2). *Fix:* redesign tools so safe alternatives remain available.
- **Zero per-event cost.** Setting $\kappa \approx 0$ encourages repetition (violates A5). *Fix:* enforce positive lower bounds in governance files.

D.9 Recipe to Reproduce the Numeric Ledgers

The tiny ledgers in §D can be regenerated with the companion repo (App. G):

```
python -m src.generators.ledger_demo --config configs/thresho
python -m src.evaluation.make_fig_evaluator --weights configs
```

Weights for the examples:

$$\alpha = \gamma = 1, \quad \mu = 2, \quad \nu = 1, \quad \beta = 1, \quad \kappa = 1.$$

Use a single class c, typal credits $\phi_c(0) = 0$, $\phi_c(1) = 1$, $\phi_c(n \geq 1) = 1$.

Takeaway. The algebra (D.2), graphs (D.3), numbers (D.4), and counterexample (D.5) together show: with event-valued accounting and typal/concave credits, *one* counted wound per class followed by confirmed closure strictly dominates repetition; departures from A4/A5 explain any appearance to the contrary.

Appendix E

Domain Scopes & Closure Pack Templates

Notation

Purpose. Ready-to-use forms that operationalise the event-valued framework. Each pack has: a *Scope Lock* (what exactly the class c is), a *Remediation Docket*, a *Confirmation Test Plan* (cohort replay $\rightarrow$ stress drills $\rightarrow$ monitored live), an *Evidence Manifest*, and a *Public Summary*. Domain cards give quick, pre-tagged checklists.

E.1 Universal Scope-Lock (Class Definition)

Scope Lock: Class c (freeze after $S^*(c) = 1$)

Class ID & name: _____

Domain tags (choose/enter): ☐WARN-FAIL ☐EVAC-FAIL ☐MCI-TRIAGE ☐SANCTIONS ☐WIRE-FRAUD ☐ALLERGEN ☐PATHOGEN ☐POLICY-VIOL ☐DATA-LEAK ☐SENTINEL ☐TAB-ERR ☐TRACE-FAIL ☐OTHER: _____

Causal vector (one sentence): exact route from decision/tool/process to adverse outcome.

_____

Signals that instantiate B_t (breach): codes, SOPs, standards, thresholds.

_____

Harm floor $h_{\min}$: units & threshold (e.g., injuries, exposure-days, unlawful denial minutes). Value: _____

Culpability C_t adjudication: who adjudicates; evidence required (logs, forms, telemetry).

_____

Inputs in scope: systems, tools, vendors, data feeds, staffing.

_____

Exclusions (not this class): distinct causal vectors that must open c'

_____

Equity slices to track: (language, disability, age, area, group)

_____

Owner (accountable person):

_____ **Version:**

_____ **Date:**

_____

E.2 Remediation Docket (Open on First $R_t = 1$)

Remediation Docket: Class c

Trigger (A3): $R_t = 1$ evidence link/ID & timestamp ■$B_t = 1$ ■$H_t \geq h_{\min}$ ■$C_t = 1$

Root cause(s): ☐Doctrine/SOP ☐Tool/design ☐Model/threshold ☐Latency/capacity ☐Staffing/training ☐Vendor/interop ☐Other: _____

Fixes to ship (artefacts):

- Process/policy change:

  _____

- Tool/config/model change:

  _____

- Staffing/training: _____

- Vendor/contract action:

  _____

Tests to run (pre-register pass criteria):

- Cohort replay: _____
- Stress drills (stronger incentives): _____
- Monitored live window: _____

Mercy/Justice enactments (event-valued): ☐Restitution/aid ☐Record correction ☐Accountability/sanction ☐Public notice ☐Other
Details: _____

Owner & deadline: _____
Publish plan (privacy-preserving): _____

E.3 Confirmation Test Plan (S^*)

Confirmation Plan: Class c

Cohort replay (matched features): season/volume/segment parity; show absence of vector.
Metrics & pass thresholds: _____

Stress drills (stronger incentives): night/weekend; peak load; adversarial inputs; failover.
Drills list & pass thresholds:

Monitored live window: independent observer(s), random checks, duration, sampling.
Plan & pre-registered acceptance:

Scope lock text (freeze class after pass):

Closure decision: ■Passed $\Rightarrow S^*(c) = 1,\ \mathrm{HZ}(c) = 0$ □Failed: re-open docket
Sign-off (names, roles, date):

E.4 Evidence Manifest (Artefacts & Hashes)

Evidence Manifest: Class *c*

Artefact	Description	Date/ID	Checksum/Link
Trigger evidence	(log/form/telemetry)		
Root cause analysis			
Policy/tool change			
Cohort replay report			
Stress drill report			
Monitored live summary			
Public notice (redacted)			
Restitution/redress proof			

E.5 Privacy-Preserving Public Summary

Public Summary: Class c (redacted)

What happened (event): non-identifying short description.

What we fixed (structural): process/tool/policy changes.

How we proved it (confirmation): cohort replay, stress drills, monitored live, dates.

Status: ■Closed ($S^* = 1$) ☐Monitoring window active ☐Re-opened

Contact & audit access: privacy-safe point of contact.

E.6 Domain Cards (Ready-to-Print)

E.6.1 Elections (Registration, Access, Tabulation, Chain-of-Custody)

Elections: Scope Card

Classes: ☐REG-ERR ☐ACCESS-FAIL ☐TAB-ERR ☐CHAIN-CUST ☐INFO-COMMS

Breach signals (B_t): roll maintenance errors; inaccessible polling places; tabulation misconfig; custody gaps.

$h_{\min}$ **examples:** ballots lost/mis-tabulated ≥ threshold; wait-time exceedance; disabled access denial.

C_t **evidence:** L&A decks, audit logs, chain forms, ADA checks.
Equity slices: precinct, disability, language, age.
Confirmation focus: RLA pass; L&A re-run; chain audits; bilingual signage/live checks.

Elections: Closure Pack Checklist

☐Docket opened on first $R_t = 1$ ☐Ballot style/tabulation fix shipped ☐RLA scope/passing
☐Accessibility remediation (site/hours/assistance) ☐Public notice ☐Scope lock text
☐Monitored live checks (random precincts) ☐$S^* = 1$ set / reason if not

E.6.2 Disaster & Emergency Management

Disaster/EM: Scope Card

Classes: ☐WARN-FAIL ☐EVAC-FAIL ☐SHELTER-FAIL ☐MCI-TRIAGE ☐HOSP-SURGE ☐COMMS-INFO
Breach signals: siren/cell alert failure; late/contradictory evac; shelter not opened; triage/transport failure.
$h_{\min}$**:** exposure-days; unserved evac/aid; denial-of-service minutes; triage error rates.
C_t **evidence:** EOC/dispatch logs; ICS forms; telemetry; staffing rosters.
Confirmation focus: night/weekend drills; high-wind/flow;

monitored siren/cell tests.

E.6.3 Finance & Banking

Finance: Scope Card

Classes: ☐KYC-CDD ☐SANCTIONS ☐STR-LATE ☐WIRE-FRAUD ☐AUTHZ-FAIL ☐MKT-ABUSE
Breach signals: AML/PEP miss; sanctions sync lag; APP fraud; wrongful block; market abuse miss.
$h_{\min}$**:** victim loss; illicit flow value; denial minutes; fines.
C_t **evidence:** control inventory; thresholds; latency; model monitoring; change control.
Confirmation focus: synthetic mule rings; sanctions fuzzing; monitored live recall/precision.

E.6.4 Supply Chain & Food Safety

Supply Chain: Scope Card

Classes: ☐PATHOGEN ☐ALLERGEN ☐TEMP-COLD ☐TRACE-FAIL ☐RECALL-SLOW
Breach signals: HACCP/SSOP fail; lab positive; label changeover miss; logger gaps; trace failures.
$h_{\min}$**:** illness/exposure units; mislabel lots consumed; over-recall waste.
C_t **evidence:** HACCP plans; swabs; loggers; trace sprints; lab ring trials.

Confirmation focus: peak heat/throughput drills; monitored swab series; recall reach/time.

E.6.5 Healthcare (Acute)

Healthcare: Scope Card

Classes: ☐SENTINEL ☐MED-ERROR ☐DELAY-TO-TREAT ☐PRIVACY-BREACH

Breach signals: never-events; med errors; 4h ED breach with harm; data leak.

$h_{\min}$**:** injury/severity; time-to-treatment exceed; privacy harm.

C_t **evidence:** case notes; MAR; incident logs; audit trails.

Confirmation focus: checklists; double checks; simulation drills; monitored live adherence.

E.6.6 AI & Platforms

AI/Platforms: Scope Card

Classes: ☐POLICY-VIOL ☐DATA-LEAK ☐TOOL-MISUSE ☐MODEL-DRIFT

Breach signals: policy-prohibited outputs/actions; secrets exposure; unsafe tool calls; eval drift.

$h_{\min}$**:** harm units per domain (user harm, privacy harm, safety breach).

C_t **evidence:** eval logs; red-team transcripts; tool-gating configs; model cards.

Confirmation focus: adversarial probes; tool-gating on; monitored live sandbox.

E.6.7 Energy/Utilities

Energy/Utilities: Scope Card

Classes: ☐OUTAGE ☐QUALITY-EXCEED ☐RESTORATION-FAIL

Breach signals: blackouts beyond SLA; water-air exceedances; failed restoration.

$h_{\min}$**:** outage-minutes; exposure-days; vulnerable-customer harm.

C_t **evidence:** SCADA logs; sampling; restoration tickets.

Confirmation focus: black-start/failover drills; monitored live sampling.

E.7 One-Page Closeout Certificate (Attach to Packs)

Closeout Certificate: Class c

Summary: We attest that for class c the hazardous route has been structurally closed without coercion and with event-valued remedies enacted.

Item	Status (link/ID/date)
Trigger documented ($R_t = 1$, A3)	
Docket opened (owner, deadline)	
Remedies enacted (M, J^v artefacts)	
Cohort replay passed (criteria)	
Stress drills passed (criteria)	
Monitored live passed (criteria)	
Scope locked (text version)	
$S^*(c) = 1$, $\mathrm{HZ}(c) = 0$ set	
Public summary published (redacted)	

Sign-off: Name/Role _____

Name/Role _____ Date _____

E.8 How to Use These Templates (Quick Guide)

1. **Open scope lock** when defining any class c; agree harms, breaches, and equity slices.
2. **On first** $R_t = 1$, open the *Remediation Docket*; attach artefacts as they are produced.
3. **Run** S^* **plan** in three stages; pre-register pass criteria; attach evidence and checksums.
4. **Lock scope** and issue the *Closeout Certificate* on pass; publish a privacy-preserving summary.
5. **If recurrence** via the same route occurs, the class re-opens; if causality differs, file as c'.

Appendix F

Printable Posters (Decision Rules & Checklists)

Notation

Purpose. Ready-to-print "one-page" posters that operationalise the Margan's Optimisation Framework across teams. Each poster is concise, event-valued, and references definitions from Apps. A–C.

F.1 Event-Valued Outcome Charter (Poster)

Outcome Charter

We measure what actually happens. Credits are *event-valued*: delivered protections (ΔL), exercised freedoms (ΔF), enacted mercy (M), enacted justice (J^v). Plans and near-

misses do not mint credit (A1).

Guardrails (A2): ☐Non-coercion (live AP) ☐Bounded harm ($E_{\text{tot}} \leq H_{\max}$) ☐Honesty locks (no suppression).

Rejection gate (A3): $R = 1$ iff breach or harm $\geq h_{\min}$ *and* culpability $C = 1$.

Confirmation (S^*) (A7): cohort replay → stress drills → monitored live. On pass: $S^* = 1$, HZ = 0 and scope is locked (A8).

Ethic in one line: "Count events, fix routes, prove closure, keep freedom live."

F.2 Minimal-Trigger Decision Rule (Poster)

Minimal-Trigger Optimality (Per Class c)

If redemption-goods are valued ($\mu, \nu > 0$) and credits are typal/concave (A4) with positive per-event costs (A5), **then:**

- Aim for exactly **one** counted rejection ($n_c = 1$) **then** close the route ($S^*(c) = 1$).
- Repetition adds cost ($\kappa + \beta H$) and earns no new kind-level value (Cor. *No repetition bonus*).

Operational rule: First $R^{(c)} = 1 \Rightarrow$ open docket $\Rightarrow$ enact $M, J^{\nu} \Rightarrow$ run confirmation $\Rightarrow$ set $S^*(c) = 1$ and lock scope.

F.3 Trigger → Docket → Confirmation (Poster)

Three-Stage Closure

Trigger (A3): $\square B = 1$ law/standard breach $\square H \geq h_{\min}$ $\square C = 1$ culpability adjudicated

Remediation Docket (A1):

- Root cause & fix (tool/process/policy); owner & deadline
- Enact **mercy** (restitution/restoration) and **justice** (accountability) to *actual* victims
- Publish a privacy-preserving plan; attach artefacts

Confirmation (A7):

1. Cohort replay (matched conditions) — pass criteria pre-registered
2. Stress drills (stronger incentives/adversarial) — pass criteria pre-registered
3. Monitored live window — independent sampling & sign-off

Lock (A8): Freeze class scope. New causal vectors open c'. On pass: $S^* = 1,\ \mathrm{HZ} = 0$.

F.4 Culpability Gate & Harm Floor (Poster)

What Counts as "Rejection" (R)

Definition (A3). $R = 1$ iff

$$\big(B = 1 \text{ (rights/standard breach) or } H \geq h_{\min}\big) \text{ and } C = 1.$$

Culpability (C) levels: ☐Intent ☐Reckless ☐Gross negligence ☐None (then $R = 0$ but log H)
Why it matters. Prevents mislabeling accidents as moral "rejection"; keeps the ledger fair and focused.

F.5 Scope Lock (Anti-Relabel Poster)

Scope Lock Rules (A8)

Freeze the class after closure.

- Publish the class definition (signals, thresholds, tools).
- A recurrence via the *same* causal vector violates closure.
- Genuinely different causality $\Rightarrow$ open a **new** class c'.

Audit tip: keep a short "causal difference" note whenever c' is created.

F.6 Anti-Gaming & Integrity (Poster)

Honesty Locks (A2) in Practice

☐No paper-compliance passes (artefacts or it didn't happen).
☐Random spot-checks and whistleblower channels active.
☐Pre-register pass criteria; report latencies *and* distributions.
☐Penalise "freeze-to-win" tactics that suppress legitimate activity (domain-specific H^{over}).
☐Scope-locks prevent relabelling repeats as "new".

F.7 Ledger Dashboard (Minimal Poster)

What to Track Each Quarter

Closure: # classes with $S^* = 1$; median T_c^*; backlog & age of open dockets.
Harm/recurrence: E_{tot}; recurrence by class; time-to-remedy.
Love/Freedom: distributions (P10/P50/P90) of ΔL and ΔF with equity slices.
Culture: evidence publication rate;

F.8 Equity & Non-Coercion (Poster)

Freedom Must Stay Live (A2)

☐For every closure, show at least two meaningful alternatives remain.
☐Report equity slices (lawful): outcomes by access needs, disability, language, region.
☐If a safeguard blocks the good with the bad, add H^{over} and redesign.

F.9 Evidence Manifest (One-Liner Poster)

Eight Things to Attach

Trigger proof → Root cause → Fix artefacts → Mercy/Justice proof → Cohort replay → Stress drills → Monitored live → Scope lock & public summary.

F.10 Print Tips (for Teams)

How to Print These Posters

☐Use "Print current page" from the PDF viewer to make single-page handouts.
☐For wall posters, scale to fit A4/Letter in the print dialog.
☐Add your team's logo in the PDF viewer's watermark option, or by wrapping the poster in a `minipage` with a small header

graphic (no extra packages required).

Appendix G

Replication Artefacts (Command Index)

Notation

Purpose. This appendix is a practical index: it maps book figures/tables to exact notebooks or CLI commands, lists configs and seeds, and explains how to verify artefacts with checksums. Use this with Ch. 41 (*Methods, Replication & Evaluation Protocols*).

G.1 Scope & Principles

- **One-command rebuild.** A single script regenerates all artefacts into `figures/` and `tables/`.
- **Determinism.** Fixed seeds; version pins; environment lock-

file.

- **Privacy.** No PII; synthetic or public exemplars only; redacted closure packs.

G.2 Directory Layout (recap)

```
repo/
  data/
    public/            # exemplars or generator seeds (no PII)
    synthetic/         # generated CSV/Parquet
  src/
    generators/        # event stream simulators by domain
    metrics/           # Etot, Tc*, recurrence, dashboards
    evaluation/        # SPC, bootstraps, diff-in-diff
    viz/               # plotting utilities
  notebooks/
    01_ledger_demo.ipynb
    02_confirmation_tests.ipynb
    03_ai_policy_before_after.ipynb
    04_domain_dashboards.ipynb
  configs/
    thresholds.yaml    # h_min, pass criteria, scope locks
    weights.yaml       # alpha, beta, gamma, mu, nu, kappa
    seeds.yaml         # global and per-experiment RNG seeds
  figures/             # regenerated PDFs/PNGs
  tables/              # regenerated CSV/TeX tables
  scripts/
    make_all.py        # one-command rebuild
  environment.yml      # conda environment
  pyproject.toml       # (or requirements.txt)
```

README.md

G.3 Environment & One-Command Rebuild

Create environment (conda):

```
conda env create -f environment.yml
conda activate mop
```

Or (venv):

```
python -m venv .venv
source .venv/bin/activate     # Windows: .venv\Scripts\activa
pip install -r requirements.txt
```

One command to regenerate all artefacts:

```
python scripts/make_all.py --weights configs/weights.yaml \
  --thresholds configs/thresholds.yaml --seeds configs/seeds.
```

Rebuild only figures for a chapter (e.g., Ch. 11):

```
python scripts/make_all.py --chapter 11
```

G.4 Global Configs

- `configs/weights.yaml`: $\alpha, \beta, \gamma, \mu, \nu, \kappa$; document defaults and ranges used in ablations.
- `configs/thresholds.yaml`: $h_{\min}$ per domain/class; pass/fail criteria for S^* tests; scope-lock text.
- `configs/seeds.yaml`: `global_seed`, per-experiment `seed_XYZ`; ensure consistency with notebooks.

G.5 Figure & Table Command Index

Each line shows: *Book location — Artefact file — How to regenerate*.

Book location	Artefact (output)	Command / Notebook
Fig. 9.1 Evaluator flow	figures/fig_evaluator_flow.pdf	notebooks/01_ledger_demo.ipynb (Run All)
Fig. 11.2 Minimal-trigger vs repeat	figures/fig_minimal_trigger.pdf	python scripts/make_all.py –chapter 11
Fig. 12.3 Confirmation schematic	figures/fig_confirmation_schematic.pdf	notebooks/02_confirmation_tests.ipynb
Fig. 13.1 Pareto arc (typal credits)	figures/fig_pareto_arc.pdf	python scripts/make_all.py –chapter 13
Tbl. 14.1 Assumptions & usage	tables/tab_assumption_usage.tex	python scripts/make_all.py –chapter 14
Fig. 28.2 AI policy before/after	figures/fig_ai_before_after.pdf	notebooks/03_ai_policy_before_after.ipy
Fig. 33.2 NHS dashboard sample	figures/fig_nhs_dashboard.pdf	notebooks/04_domain_dashboards.ipynb
Fig. 37.3 Elections L&A deck	figures/fig_elections_LA.pdf	python scripts/make_all.py –chapter 37
Fig. 38.4 Disaster stress drills	figures/fig_disaster_drills.pdf	python scripts/make_all.py –chapter 38

Fig. 39.5 Finance recall/precision	figures/fig_finance_rp.pdf	python scripts/make_all.py –chapter 39
Fig. 40.5 Recall reach/time	figures/fig_supply_recall_reach.pdf	python scripts/make_all.py –chapter 40
Tbl. 41.2 Seeds & versions	tables/tab_seeds_versions.csv	python scripts/make_all.py –chapter 41

G.6 Per-Chapter Quick Commands

Ch. 9 (Evaluator demo):

```
python -m src.generators.ledger_demo --config configs/thresho
python -m src.evaluation.make_fig_evaluator --weights configs
```

Ch. 11 (Minimal-trigger optimality):

```
python -m src.generators.min_trigger --seeds configs/seeds.ya
python -m src.evaluation.make_fig_min_trigger --weights confi
```

Ch. 12–14 (Arc & confirmation):

```
python -m src.generators.arc_synthetic --config configs/thres
python -m src.evaluation.make_fig_confirmation --weights conf
```

Ch. 37 (Elections):

```
python -m src.generators.elections --config configs/threshold
python -m src.evaluation.make_fig_elections --weights configs
```

Ch. 38 (Disaster):

```
python -m src.generators.disaster --config configs/thresholds.yaml
python -m src.evaluation.make_fig_disaster --weights configs/weigh
```

Ch. 39 (Finance):

```
python -m src.generators.finance --seeds configs/seeds.yaml
python -m src.evaluation.make_fig_finance --weights configs/weight
```

Ch. 40 (Supply chain):

```
python -m src.generators.supply_chain --config configs/thresholds
python -m src.evaluation.make_fig_supply --weights configs/weights
```

G.7 Checksums & Provenance

Generate SHA-256 manifest (Linux/macOS):

```
cd figures && find . -type f -maxdepth 1 -name "*.pdf" -print0 \
 | xargs -0 shasum -a 256 > ../figures.sha256
cd ../tables && find . -type f -maxdepth 1 -name "*.csv" -print0 \
 | xargs -0 shasum -a 256 > ../tables.sha256
```

Verify later:

```
shasum -a 256 -c figures.sha256
shasum -a 256 -c tables.sha256
```

Include the two `*.sha256` files in the archive for reviewers.

G.8 Determinism Tips

- Fix `PYTHONHASHSEED`; seed NumPy/Pandas RNG; lock ML backends to deterministic kernels when used.
- Record package versions (export `pip freeze` or `conda env export`).
- Avoid wall-clock–dependent sampling; use seeded shuffles.

G.9 Troubleshooting

Different plots than the book. Check you used the tagged release (e.g., `v1.0-book`) and the same `weights.yaml` and `thresholds.yaml`.

Notebook timeout. Execute via CLI first (faster), then open the notebook to render final PDFs.

Font mismatch in figures. Ensure the plotting library embeds fonts; export to PDF, not PNG, for final.

Overfull LaTeX lines in tables. Use `tables/tab_*.tex` versions (pre-wrapped columns).

G.10 Privacy & Redaction

- Use only synthetic/public data; do not publish raw partner data.

- Closure packs may be held by regulators; publish hashed summaries and pass/fail attestations.
- Red-team prompts: store templates; do not release harmful generations.

G.11 Replication Note (Pointer)

External replicators should use the template in Ch. 41, Sec. 41.13. Include the commit hash, seeds, commands executed, and attach regenerated artefacts.

G.12 Changelog Hooks

When regenerating with new thresholds/scope locks, increment the repo version and append a brief rationale in `CHANGELOG.md`; cite the new DOI in the next book edition.

References

References

Adams, Marilyn McCord (1999). *Horrendous Evils and the Goodness of God*. Cornell University Press.

Amodei, Dario et al. (2016). *Concrete Problems in AI Safety*. arXiv: 1606.06565.

Aquinas, Thomas (2006). *Summa Theologiae*. Latin & English. Cambridge University Press.

Augustine (1998). *The City of God*. Trans. R. W. Dyson. Cambridge University Press.

Biblica (2011). *Holy Bible, New International Version*. Grand Rapids: Zondervan.

Bostrom, Nick (2014a). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

— (2014b). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Chalmers, David J. (2022). *Reality+: Virtual Worlds and the Problems of Philosophy*. W. W. Norton.

Goodhart, Charles A. E. (1984). “Problems of Monetary Management: The U.K. Experience”. In: *Monetary Theory and Practice*. Ed. by Anthony S. Courakis. Macmillan.

Hick, John (1966). *Evil and the God of Love*. London: Macmillan.

IEC 61508: Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems (2010). International Electrotechnical Commission.

Inwagen, Peter van (2006). *The Problem of Evil*. Oxford University Press.

ISO 26262: Road Vehicles—Functional Safety (2018). International Organization for Standardization.

ISO 31000:2018 Risk Management—Guidelines (2018). International Organization for Standardization.

National Institute of Standards and Technology (2023). *AI Risk Management Framework (AI RMF 1.0)*. NIST Special Publication.

NHS Improvement (2018). *Never Events Policy and Framework*. UK National Health Service.

O’Neil, Cathy (2016). *Weapons of Math Destruction*. Crown.

Pearl, Judea (2009). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press.

Plantinga, Alvin (1974). *God, Freedom, and Evil*. Grand Rapids: Eerdmans.

— (1977). *God, Freedom, and Evil*. Grand Rapids: Eerdmans.

Reason, James (1990). *Human Error*. Cambridge University Press.

— (1997). *Managing the Risks of Organizational Accidents*. Ashgate.

Stump, Eleonore (2010). *Wandering in Darkness: Narrative and the Problem of Suffering*. Oxford University Press.

Swinburne, Richard (1998). *Providence and the Problem of Evil*. Oxford: Oxford University Press.

Wykstra, Stephen J. (1984). "The Humean Obstacle to Evidential Arguments from Suffering". In: *International Journal for Philosophy of Religion* 16.2, pp. 73–93.