Computational Closure and the Architecture of Mind: An Information-Theoretic Foundation for Naturalized Epistemology

Abstract

Naturalized epistemology faces a fundamental question: how do physical brains, bound by thermodynamic constraints, come to grasp objective truths about reality? This paper proposes that the bridge lies in information compression. We argue that truth is not static correspondence to facts, but rather alignment with the Optimal Constraint Configuration: the compression structure that minimizes information leakage against reality's causal constraints. This represents a thermodynamic attractor for viable knowledge systems.

We distinguish two modes of pattern recognition: statistical pattern matching, which extracts correlations through frequency, and structural pattern recognition, which detects necessary relationships between mutually constraining components. This distinction explains why some knowledge requires extensive sampling (ravens are black) while other knowledge achieves validity from singular encounters (fire burns). We propose, tentatively, that consciousness may track structural pattern recognition, serving as the cognitive interface for grasping structural coherence.

Drawing on the Free Energy Principle and computational mechanics, we show how cognition compresses information through Markov blankets (statistical boundaries) that achieve computational closure when macro-level dynamics become causally autonomous from micro-level substrates. Our contribution is conceptual integration across information theory, cognitive science, and epistemology, showing how these frameworks illuminate each other when properly connected.

The framework remains speculative and does not solve the hard problem of consciousness. However, it identifies functional correlates that narrow the explanatory target and offers a naturalistic account of how physical constraints give rise to epistemic norms. Alternative explanations remain viable, and the account requires further development and testing.

1. Introduction

This paper is a work of synthesis rather than empirical discovery. It presents no new experimental data in neuroscience, nor new theorems in thermodynamics or information theory. It aligns existing, well-validated concepts from information geometry (computational closure), cognitive science (predictive processing), and epistemology (Quinean holism) to resolve friction between them. Our contribution is architectural alignment: showing how these frameworks illuminate each other when properly connected.

Contemporary theories of mind and knowledge largely talk past one another. Cognitive science increasingly describes the brain as a prediction engine minimizing free energy to persist in its environment (Friston 2010). This is a fundamentally thermodynamic process. Meanwhile, naturalized epistemology describes knowledge as a web of beliefs justified by its coherence and pragmatic success (Quine 1960; Sinclair 2007). It makes little reference to the underlying mechanics. This leaves an explanatory gap: how do the physical constraints on an organism's existence give rise to the normative structures of justification and objective truth? (BonJour 1985; Olsson 2005)

While other attempts to bridge this gap exist (e.g., evolutionary epistemology emphasizing adaptation or social epistemology focusing on testimony), this paper tentatively proposes that the bridge may lie in information compression. By treating concepts not as abstract representations but as *Markov blankets* (statistical boundaries that separate internal from external states) that achieve *computational closure* (a hypothesized state where the macro-level is causally self-contained), we may be able to trace a continuous line from the thermodynamics of living systems to the structure of belief. We suggest that the drive to minimize prediction error could be the engine that shapes both cognitive architecture and epistemic norms, though this remains a theoretical proposal requiring empirical validation.

Methodologically, this framework adheres to what Ladyman and Ross (2007) term the *Principle of Naturalistic Closure*: we reject a priori intuition as a guide to metaphysics and instead require that epistemological claims unify the special sciences with fundamental physics without contradiction. The generative priors and Markov blankets described herein are not merely psychological heuristics but are motivated by the convergence of thermodynamics and information theory, a convergence that grounds cognition in the physical constraints any bounded system must satisfy.

Our central thesis, which remains speculative, is that consciousness may function as the user interface for a specific type of information processing: structural pattern recognition. We distinguish this from the unconscious, frequency-based processing of statistical regularities. This distinction potentially explains why some knowledge can be acquired from a single instance and may provide a functional basis for the phenomenology of understanding. It also offers a prospective diagnostic for the limitations of current artificial intelligence, though this application requires further investigation.

The argument proceeds as follows. First, we establish the information-theoretic foundations of bounded systems, defending the Free Energy Principle as a constitutive condition (Section 2). Second, we detail how Markov Blankets and computational closure create emergent causal levels (Section 3). Third, we develop our theory of consciousness based on the statistical/structural distinction, separating theoretical foundations (Section 4A) from phenomenological applications (Section 4B). Fourth, we show how logic and mathematics emerge as necessary compression structures (Section 5). Fifth, we examine emergence through computational closure (Section 6). Sixth, we show how this architecture culminates in a naturalized account of truth as alignment with the optimal constraint configuration: the thermodynamic attractor for any viable knowledge system (Section 7). Finally, we explore applications to macro-epistemology and systemic brittleness (Section 8).

Conceptual Roadmap: This paper traces a continuous path from physical constraints to epistemic norms: from information theory and Markov blankets to

computational closure, consciousness, logic/mathematics, emergence, the optimal constraint configuration, and finally to applications. Each section builds systematically on the previous one, creating a unified framework that bridges thermodynamic mechanisms with normative epistemology.

2. Information as Fundamental Substrate

2.1 The Primacy of Information

All physical systems process information. A rock in sunlight absorbs photons (information input) and radiates heat (information output). Most systems, however, are informationally transparent: information flows through without creating persistent internal structure. The rock dissipates absorbed energy immediately as heat, performing no computational work. It does not build a model of sunlight patterns or predict tomorrow's weather. It simply responds mechanically to immediate inputs.

Living and cognitive systems differ fundamentally. They compress information, building internal models that predict future sensory states. Rather than immediately dissipating absorbed energy, they delay dissipation to perform computational work: encoding regularities, updating predictions, guiding action. This compression is literal (in Shannon's sense), not metaphorical: reducing surprise by encoding regularities into reusable patterns.

A Note on Entropy and Energy: Throughout this paper, we distinguish between information-theoretic quantities (Shannon entropy as average surprise, measured in bits) and thermodynamic quantities (energy required to process or erase bits). These are not identical but deeply connected. Landauer's Principle establishes that information processing has irreducible physical costs. While erasing a single bit costs negligible energy, maintaining a high-fidelity internal model against a shifting environment requires continuous processing of massive information streams. At the macro-scale, this 'information processing cost' becomes metabolic and economic overhead. The 'heat' generated by a failing belief system manifests as wasted bureaucratic effort, enforcement friction, and cognitive dissonance. High information leakage (persistent prediction error) thus implies tangible costs to any physical system. The link between epistemic brittleness and physical inefficiency is functional, not merely metaphorical.

Persistence as a bounded pattern requires information processing. What we call "things" are better understood as *real patterns* in the sense of Dennett (1991) and Ladyman and Ross (2007). Crucially, we must distinguish between material substrate (the underlying physical constituents) and causal structure (the constraints that shape outcomes). A real pattern exists not by virtue of what it is made of, but by what it does to a probability distribution. If a pattern reduces uncertainty and supports counterfactual interventions (if heating this boundary causes the interior to produce heat shock proteins), it constitutes an ontological unit. Real patterns compress information, maintaining statistical boundaries that distinguish internal from external states while achieving genuine projectibility (predictive purchase on future states).

2.2 The Free Energy Principle

Karl Friston's Free Energy Principle provides one well-developed formalization of these dynamics:

Variational Free Energy = Surprise (unpredicted sensory input) + Divergence (model complexity)

All self-organizing systems minimize free energy by: 1. **Updating beliefs** to better predict sensory input (perceptual inference) 2. **Changing the world** to match predictions (active inference) 3. **Optimizing model structure** to reduce complexity while maintaining accuracy (structural learning)

This process is not a goal-directed endeavor but a physical constraint: systems that fail to minimize free energy tend to dissipate and vanish, whereas those that succeed maintain their integrity as bounded entities. In this sense, minimizing free energy is constitutive of self-organization (Friston 2010).

The Free Energy Principle (FEP) builds on broader predictive processing frameworks in cognitive science, where brains are understood as hierarchical prediction machines constantly minimizing prediction error through bidirectional cortical processing (Clark 2013). This perspective reframes perception. It treats perception not as passive reception but as active inference, where the brain tests predictions against sensory input and revises models when mismatches occur. Neurobiologically, this is implemented through predictive coding: hierarchical brain organization where higher levels generate predictions about lower-level activity, and lower levels signal back prediction errors when inputs deviate from expectations (Friston and Kiebel 2009). This architecture provides a plausible mechanistic account of how variational free energy minimization could be realized in cortical circuits.

The FEP's scope extends beyond brain function to encompass fundamental properties of all self-organizing systems. Any system that maintains its integrity against the second law of thermodynamics must possess a statistical boundary. This boundary (a Markov blanket) separates internal from external states. It must minimize the free energy associated with that boundary (Friston et al. 2017).

Kirchhoff et al. (2018) demonstrate that living systems are characterized by Markov blankets enabling autonomous organization, and that "autonomous systems are hierarchically composed of Markov blankets of Markov blankets—all the way down to individual cells, all the way up to you and me." This hierarchical assembly occurs through adaptive active inference, distinguishing living systems (capable of inferring future states) from non-living systems exhibiting mere active inference without adaptive capacity.

This connects information-theoretic principles directly to the basic requirements for life itself. Autopoiesis, allostasis, and goal-directed behavior all emerge as consequences of free energy minimization in bounded systems.

Connection to Epistemic Brittleness: Systemic brittleness is accumulated free energy. When a knowledge system's predictions consistently fail (information leakage), it must either patch the model with ad-hoc additions, suppress disconfirming evidence through coercion, or accept falsification and revise. The brittleness metrics developed in Glenn (2025) (patch velocity, coercive overhead, model complexity, resilience reserve) measure these information-theoretic costs directly.

Reality functions as the constraint landscape that shapes belief revision. Each failed prediction generates an error signal: a mismatch between expected and actual outcomes that drives the system to update its compression. In Quine's terms, beliefs form a web where central propositions (those encoding high structural compression, like logic and mathematics) resist revision because changing them would require massive reorganization of the entire system. The brittleness metrics track these revision costs. When reality's constraints conflict with a system's compressions, accumulated prediction errors manifest as increasing patch velocity, rising model complexity, and declining resilience. The system faces a choice: revise the compression to align with constraints, or expend energy suppressing the error signals through coercion.

2.3 Dispositions as Compression Algorithms

Recent formal work provides rigorous information-theoretic foundations for this compression imperative. Dittrich and Kinne (2024) demonstrate that any system persisting under uncertainty must compress information to minimize what they term "epistemic entropy." This is the Information-Theoretic Imperative (ITI): bounded existence requires efficient compression as a constitutive condition, not merely a useful strategy. Their Compression Efficiency Principle (CEP) further shows that efficient compression mechanically selects for generative, causal models over superficial correlations. Non-causal compressions "accumulate exceptions" as novel cases fail to fit the pattern, eventually becoming unsustainable. The drive to compress is thus a physical necessity for bounded systems. This mechanism explains why systems are forced to seek computational closure: only compressions achieving genuine causal structure can sustain efficiency without catastrophic exception accumulation.

This convergence of independent formalizations, including classical information theory (Shannon 1948), thermodynamic cost analysis (Landauer 1961), and evolutionary epistemology (Campbell 1974; Bradie 1986), suggests that these principles capture genuine structural features of how bounded systems interact with constrained environments, not artifacts of any particular theoretical framework.

Returning to the Quinean foundation: a disposition to assent functions analogously to a compressed encoding of regularities.

After encountering many dogs, an organism develops a "dog-detecting" disposition: a neural pattern that fires when dog-relevant features appear. This disposition compresses thousands of observations into a single reusable pattern, predicts future behavior (minimizing surprise when encountering new dogs), enables efficient action (approach friendly dogs, avoid aggressive ones), and stores mutual information across sensory streams (visual, olfactory, auditory).

Better compression means deeper understanding. A child learning "all dogs bark" has simple but lossy compression. An ethologist understanding canine communication has complex but high-fidelity compression. The compression ratio (how much information is preserved with how few parameters) tracks what we intuitively recognize as understanding.

Quine's insight that beliefs rest on dispositions to assent finds mechanistic grounding here. For Quine, these dispositions are not static logical axioms but dynamic tendencies that shift when experience contradicts expectations. This framework makes the mechanism explicit: what Quine called a 'disposition to assent' functions computationally as a 'generative prior', a parameter in a

predictive compression system, continuously adjusted to minimize surprise. We treat this as a conceptual framework for understanding the functional role of dispositions, not as a claim that the brain literally computes these parameters in this exact way. When a prediction fails (the friendly-looking dog bites), the disposition updates, not through conscious deliberation but through the same information-theoretic process by which any learning system revises its internal model to reduce future error. The parallel to how adaptive systems adjust their parameters to improve predictions is not coincidental; Quine's naturalized epistemology anticipated what information theory later formalized.

Terminological Clarification: When we speak of an "internal model" or "compression," we are using information-theoretic shorthand for what Quine identified as a complex disposition to assent. A system "has a model of fire" not because it contains a picture of fire, but because its dispositions to act are tightly coupled to the causal constraints of combustion. Computational closure is the state where these dispositions achieve maximum predictive success with minimum adjustment. Readers who prefer behavioral or dispositional language may translate our model-talk accordingly without loss of meaning.

2.4 Information Complexes and Mutual Information

Mutual information measures how much knowing one variable tells you about another. Dispositions storing high mutual information across multiple domains form information complexes: stable attractors in the space of possible compressions.

Example: The "Fire" Complex - Visual (flames, light patterns) - Thermal (heat sensation) - Olfactory (smoke, burning) - Auditory (crackling, roaring) - Social (warnings, stories about danger) - Practical (cooking, destruction, tool-making)

These information streams share latent structure: they co-vary reliably. A disposition that compresses their mutual information (the concept "fire") achieves massive compression efficiency. This is why "fire" feels like a unified thing: it is a genuine compression joint in reality's information structure.

Not all compressions are equal. Some compressions are artifacts (rain dances cause rain), others track genuine causal structure (dry wood causes fire). Reality selects for the latter through differential brittleness.

2.5 Two Types of Patterns: Statistical Regularities and Structural Coherence

Not all patterns require the same evidential basis to be validly recognized. This distinction becomes crucial for understanding how knowledge can arise from limited or even singular encounters with phenomena.

Statistical Regularities emerge from repetition and frequency. They are discovered through observing that certain patterns recur reliably across many trials. A child learns that "dogs bark" by encountering many dogs and observing the correlation between the visual pattern (dog-shape) and the auditory pattern (barking). These regularities are fundamentally frequency-dependent; the strength of the compression relies on sample size.

Structural Regularities, by contrast, involve components that mutually constrain each other through necessary relationships. These patterns can be recognized even

in singular instances because the components aren't merely correlated but necessarily linked through causal or logical dependencies.

Example: Fire

When Robinson Crusoe first encounters fire, he doesn't need hundreds of observations to form the valid belief that "fire produces heat." The relationship between combustion, heat, and light isn't merely a statistical correlation but a thermodynamic necessity. The components constrain each other: - Combustion releases energy - Energy manifests as heat and light - Heat propagates to nearby objects - The process requires fuel and oxygen

These aren't separate facts that happen to co-occur; they're aspects of a unified causal process. A mind encountering this pattern even once can recognize its structural integrity: the internal coherence that makes it a genuine compression joint rather than an accidental correlation.

Contrast with Pure Statistical Learning: - "Hot stoves burn skin" (structural: recognized from single encounter, thermodynamic necessity) - "Dogs bark at strangers" (statistical: requires multiple observations, behavioral tendency) - "$F = ma$" (structural: mathematical necessity once the concepts are understood) - "Swans are white" (statistical: inductively generalized from frequency, famously failed)

Recent work in phase epistemology provides formal treatment of this distinction. Ayvazov (2025) distinguishes between classical probability and what he terms "improbabilistic coherence." Classical probability involves frequency-based likelihood, while improbabilistic coherence refers to structural integrity that exists independent of repetition. He defines this as "the generative condition for epistemic emergence." While Ayvazov proposes a speculative quantum-mechanical formalism for this distinction, we employ it here purely as an epistemic category without committing to his physical interpretation.

Implications for Information Compression: - Statistical compressions require large ensembles to stabilize (high sample complexity) - Structural compressions can achieve validity from minimal data when the pattern exhibits internal constraint - The brain appears capable of detecting both types, but conscious reasoning particularly engages with structural patterns - Innovation often involves recognizing structural coherence before statistical validation

Structural patterns are not subjectively imposed but constrained by reality. You cannot validly infer that "heat flows spontaneously from ice to hand" from a single encounter, because thermodynamics forbids this relationship (violating the Clausius statement of the Second Law). The structural constraints are objective, even if recognizable from limited data.

Empirical work in cognitive science supports this distinction. Statistical learning operates implicitly through mere exposure to input patterns, extracting regularities from repeated encounters (Aslin and Newport 2012). However, the same mechanisms that enable learning from frequency distributions also support generalization to novel instances when structural relationships are detected, suggesting a unified learning system capable of both statistical pattern matching and structural inference.

This distinction becomes essential for understanding how notions (proto-Generative Priors) can form before extensive empirical testing, and why some singular experiences carry immediate epistemic weight while others require statistical accumulation.

2.5.1 Addressing the Measurement Challenge

While the distinction between statistical and structural regularities provides conceptual clarity, measuring these patterns in practice presents challenges. We propose a "Constrained Interpretation" methodology that manages hermeneutic circularity through physical-biological anchors, comparative-diachronic analysis, and convergent evidence. This approach provides pragmatic objectivity sufficient for comparative assessment, though it does not eliminate interpretive challenges entirely.

The methodology employs three complementary strategies: - **Physical-Biological Anchors**: Ground interpretations in well-established physical and biological constraints (thermodynamics, evolutionary biology) that provide objective reference points - **Comparative-Diachronic Analysis**: Track how interpretations evolve over time and across cultures, identifying patterns of convergence that suggest structural rather than statistical validity - **Convergent Evidence**: Require multiple independent lines of evidence to support structural claims, reducing reliance on any single interpretive framework

These methods provide pragmatic objectivity sufficient for comparative epistemic assessment, though they do not eliminate interpretive challenges.

2.5.2 The Interventional Diagnostic

How does a cognitive system distinguish a robust structural constraint from a brittle statistical regularity? We propose the mechanism lies in the capacity for interventional counterfactuals (formally, the distinction between $P(Y \mid X)$ and $P(Y \mid do(X))$ (Pearl 2000)).

Statistical patterns capture passive observations: seeing X predicts Y (conditioning on X). Structural patterns capture interventional truths: *causing* X necessitates Y (intervening on X). This latter operation, the "dictator" or "do-operator," involves "graph surgery"; it severs the incoming causal arrows to X to isolate its downstream effects from confounding variables.

Consider a barometer. Observing the needle drop predicts rain ($P(Rain \mid NeedleDrop)$ is high). But manually forcing the needle down (intervention, $do(NeedleDrop)$) breaks the correlation; it yields no rain ($P(Rain \mid do(NeedleDrop))$ is low). The pattern was merely statistical, dependent on a common hidden cause (atmospheric pressure). By contrast, igniting fuel is structural: causing the fire ($do(Fire)$) invariably produces heat. The relationship is invariant under intervention.

Phenomenologically, the do-operator captures the cognitive difference between watching and poking. When a child switches from merely observing a spinning top to spinning it themselves, they move from statistical observation to structural intervention. The transition is not just behavioral but epistemic: passive observation reveals correlations, while active manipulation reveals causal structure. This is why hands-on learning often produces deeper understanding than passive instruction. The child who spins the top grasps its dynamics in a way the mere observer does not.

Consciousness generally tracks this distinction. While not all systems that represent causal structure are conscious, conscious systems must be capable of representing this difference between seeing and doing. The "Aha!" moment of understanding corresponds to the validation of a pattern under simulated intervention, confirming that the relationship holds not just in the observed distribution, but in the counterfactual space of actions.

The Do-operator is the test for whether a model has captured causal structure; computational closure occurs when that causal structure can be represented entirely at the macro-level. A pattern that disappears under intervention is merely statistical; a pattern that survives intervention is causal; a pattern whose causal structure closes at its own boundary is a real whole.

2.5.3 The Cognitive Signature of Structural Recognition

The interventional diagnostic provides a retrospective test for structural patterns, but cognitive systems must make real-time discriminations before extensive testing. How does a mind encountering fire for the first time recognize it as a structural pattern worthy of immediate trust, while treating raven coloration as requiring further sampling?

We propose that structural patterns exhibit a characteristic cognitive signature: they generate bidirectional expectations. Recognizing that fire produces heat simultaneously generates expectations that heat sources may involve combustion, that removing fuel will extinguish flame, that nearby objects will warm, that moisture will resist ignition. The components constrain each other mutually. This bidirectionality (the sense that elements mutually necessitate rather than merely predict) may be the functional marker triggering conscious engagement.

Statistical patterns, by contrast, generate unidirectional expectations. Observing a raven generates the prediction "probably black," but observing blackness generates no particular expectation about ravens. The pattern runs one way. There is correlation without mutual constraint.

This distinction is not merely subjective. The bidirectional constraint structure reflects genuine differences in the underlying causal topology. Structural patterns exhibit what graph theory calls "dense connectivity": multiple causal paths linking components in reciprocal relationships. Statistical patterns exhibit "sparse connectivity": few or one-way linkages vulnerable to confounding. The phenomenological difference (the "felt necessity" of structural patterns versus the tentative quality of statistical generalizations) may track this topological difference.

Implications for knowledge acquisition: This explains why some singular experiences carry immediate epistemic weight while others require accumulation. When multimodal information streams converge on mutually constraining relationships (visual flame + thermal radiation + auditory crackling + olfactory combustion products), the brain detects structural coherence. The pattern "explains itself" through internal constraint relationships, requiring minimal external validation. When information streams merely correlate without mutual constraint (black appearance + raven shape), the brain requires extensive sampling to distinguish genuine pattern from accident.

This account remains functionalist rather than mechanistic. We identify what successful structural recognition must accomplish (detecting bidirectional constraint networks) and what phenomenological markers appear to track this detection (felt necessity, immediate validity, conscious awareness). We do not specify neural implementation. Whether this involves specific cortical circuits, particular neurotransmitter dynamics, or distributed network properties remains an empirical question for cognitive neuroscience.

2.6 The Framework as Conceptual Scaffolding

We must clarify the epistemic status of the information-theoretic language employed throughout this paper.

Metaphor or Mechanism? Information theory provides conceptual scaffolding for understanding cognitive and epistemic processes, but we need not claim the brain literally computes Shannon entropy or that generative priors are implemented as explicit ε-machines. The framework's value is primarily conceptual: it captures functional relationships and constraints that appear to govern how knowledge systems operate, regardless of specific implementation details.

Analogy from Economics: Consider how economics productively uses "utility maximization" to model decision-making without claiming that neurons actually calculate utility functions. The model captures something real about choice behavior under constraints, even if the mechanistic implementation differs from the formal apparatus. Similarly, our information-theoretic framework captures something real about how compressions succeed or fail, how boundaries form and persist, and how systems minimize prediction error, whether or not these processes literally instantiate Shannon's equations.

The Philosophical vs. Empirical Claims: We can distinguish three levels of commitment:

1. **Weak Claim (Conceptual):** Information-theoretic language provides a coherent framework for thinking about dispositions, predicates, and truth. It clarifies what success and failure look like for knowledge systems.
2. **Moderate Claim (Functional):** Cognitive and epistemic systems behave AS IF they are minimizing information-theoretic costs. The functional constraints we describe (computational closure, information leakage, brittleness accumulation) genuinely constrain which belief systems persist.
3. **Strong Claim (Mechanistic):** Brains literally implement variational free energy minimization; generative priors are actually encoded as Markov blankets in neural tissue; the optimal constraint configuration exists as a definite information structure.

Our Position: This paper primarily makes claims at levels 1 and 2. Whether the strong mechanistic claims (level 3) are literally true remains an empirical question for neuroscience and cognitive science. Our philosophical insights about generative priors, compression, and truth don't depend on resolving this empirical question. Even if the brain's actual mechanisms differ significantly from Free Energy minimization, the functional analysis of what makes a predicate successful (low brittleness, high compression, computational closure) retains its philosophical force.

Acknowledging Fundamental Critiques: Recent work has challenged the FEP's scientific status directly. Mangalam (2025) argues that the FEP operates as an unfalsifiable pseudo-theory. Its high level of abstraction and mathematical formalism make it difficult to test empirically. It may not generate novel predictions that simpler theories cannot explain.

The Strong Pivot: This critique deserves a direct response, and we can offer one that strengthens rather than weakens our philosophical framework. We accept the characterization but reject the implication that it lacks explanatory power. Even if Mangalam is correct that the FEP describes necessary consequences of being a bounded system rather than contingent empirical laws, this makes it precisely the right foundation for our philosophical project.

Clarifying Our Use: We acknowledge that the FEP's generality makes it difficult to falsify as empirical neuroscience. However, our use is more constrained than the broad formulation Mangalam critiques. We apply FEP specifically to bounded systems possessing Markov blankets that successfully maintain themselves over time. This restricts the domain to systems where the principle has empirical bite: systems that fail to minimize free energy relative to their environment dissipate and cease to exist as distinct entities. While this may be "tautological" in the sense that persistent systems are those that persist, it provides substantive constraints on what architectures enable persistence. We treat FEP as organizing principle and conceptual scaffolding, not as a falsifiable hypothesis about neural implementation details. This methodological choice makes our framework more defensible precisely because it doesn't depend on resolving ongoing debates about cortical microcircuitry or exact Bayesian computation.

The FEP as Scale-Invariant Constitutive Condition: Consider analogous principles that are technically tautological yet explanatorily powerful. Evolution by Natural Selection can be reduced to "survivors survive": organisms that persist are those whose traits enabled persistence. This is tautological in structure yet profoundly explanatory because it identifies the necessary architecture of any system exhibiting heritable variation under selection pressure. Similarly, the Principle of Least Action in physics describes what any dynamical trajectory must satisfy, not contingent facts about particular systems. Logic itself is constitutively necessary: the law of non-contradiction follows from the structure of coherent reasoning itself.

These principles cannot be empirically falsified because they describe preconditions for their respective domains. Yet this status does not make them vacuous. Rather, they provide what we might call scale-invariant constitutive conditions: the necessary grammar for their domains. The FEP occupies this same structural role for self-organizing systems. If it describes necessary constraints that any bounded system must satisfy to persist against entropy (maintaining a statistical boundary requires minimizing divergence between internal model and environmental structure), then it functions as the grammar of viable self-organization.

Why Constitutive Principles Are Not Vacuous: The objection that "survivors survive" is explanatorily empty confuses logical form with explanatory content. Consider what evolutionary theory actually delivers: it explains why eyes evolved independently across dozens of lineages (optical physics constrains viable light-detection architectures), why certain body plans recur across unrelated taxa (biomechanical constraints limit viable solutions), why antibiotic resistance emerges predictably (mutation-selection dynamics under specific pressures). The tautological core ("survivors survive") generates these substantive predictions because it identifies the *selection mechanism* that filters the space of possible configurations.

Similarly, the FEP's constitutive status generates non-trivial constraints. If bounded systems must minimize prediction error to persist, then: (1) hierarchical predictive architectures should be ubiquitous across viable cognitive systems (they are); (2) systems should exhibit characteristic failure modes when prediction error accumulates faster than revision can accommodate (they do; this is precisely what brittleness metrics track); (3) the compression landscape should exhibit convergent structure across independent systems facing similar constraints (it does; mathematics, logic, and basic physical ontologies are discovered independently across cultures). The principle is "tautological" in that it describes

what persistence requires, but the architectural constraints it generates are substantive and falsifiable. A world where cognitive systems succeeded through random response rather than predictive modeling would falsify the framework's implications, even if the constitutive principle itself remained logically necessary.

The vacuity objection proves too much. By its logic, thermodynamics is vacuous ("systems that dissipate energy dissipate energy"), logic is vacuous ("coherent reasoning is coherent"), and mathematics is vacuous ("valid proofs are valid"). These domains are constitutively structured by principles that cannot be violated without exiting the domain entirely. This does not make them uninformative; it makes them foundational. The FEP, on our reading, belongs to this class: not a contingent hypothesis about what happens, but a constitutive constraint on what *can* happen for bounded systems.

Crucially, we distinguish the principle from its implementation. While specific mechanistic claims (predictive coding in cortical circuits, exact Bayesian updating in neurons) are empirical hypotheses subject to falsification, the principle itself describes the constitutive condition for identifying a system as distinct from its environment. This framework utilizes the FEP in this structural, constitutive sense: as the necessary architecture of bounded existence, not as a contingent empirical claim about neural mechanisms.

For our framework, this is ideal. We are not claiming that specific brain mechanisms literally compute variational free energy or that neurons perform Bayesian updates. We are claiming that successful knowledge systems (whatever their implementation) must satisfy functional constraints analogous to those the FEP formalizes: they must compress information about their environment, achieve computational closure to form stable boundaries, and face selection pressure based on how well their compressions align with reality's constraint structure.

If the FEP is tautological, it describes structural necessities rather than contingent mechanisms. Generative priors, on this reading, are linguistic handles for configurations that satisfy these necessary constraints. The optimal constraint configuration represents the pattern that necessarily crystallizes when bounded systems optimize their compressions under thermodynamic pressure. Brittleness measures departures from these structural necessities.

The philosophical insight remains: even if Mangalam's critique succeeds as a challenge to FEP's status as empirical neuroscience, it inadvertently validates its use as structural epistemology. Tautologies, when they capture genuine necessities, provide the most reliable foundations. We accept the critique and pivot: the FEP's value for this framework lies precisely in describing what any viable knowledge system must do, not in predicting what neurons happen to do.

Even if future research supersedes the Free Energy Principle as empirical neuroscience, the philosophical framework developed here would transfer to any replacement formalization. Our claims about compression, computational closure, and the optimal constraint configuration follow from basic constraints on bounded systems navigating finite resources, not from FEP's specific mathematical apparatus. What matters is that some formalization captures these functional necessities; the FEP provides a currently well-developed vocabulary for articulating them.

Connecting Functional Constraints to Normative Claims: Readers may wonder how later sections derive substantive claims about objective truth (Section 6) and ethics (Section 8) from what we've framed as "conceptual

scaffolding." The answer: these claims rely on level 2 (functional) constraints, not level 3 (mechanistic) implementation. The argument is not "brains compute Shannon entropy, therefore truth exists," but rather "whatever systems successfully compress reality must satisfy certain functional constraints (computational closure, minimal information leakage, strong lumpability), and these constraints determine which predicates persist." The Optimal Constraint Configuration is "objective" not because it exists as a Platonic form or neural structure, but because the functional requirements for successful compression are determined by reality's constraint structure, not by our beliefs about them. Similarly, the claim that coercion generates brittleness doesn't require literal free energy calculations; it requires only that systems refusing to model agents' autonomous responses must bear higher coordination costs. The normative force comes from functional necessity, not mechanistic implementation.

Operationalizing Brittleness: The concept of systemic brittleness plays a central role in this framework as the diagnostic tool for assessing knowledge system health. While this paper maintains focus on the philosophical foundations and conceptual architecture, detailed operationalization of brittleness diagnostics has been developed elsewhere (Glenn 2025). That work develops specific conceptual lenses for diagnosing brittleness across different domains: patch velocity $P(t)$ measuring the ratio of anomaly-resolution to novel-prediction work; coercion ratio $C(t)$ measuring security overhead versus productive investment; model complexity $M(t)$ tracking parameter growth against marginal performance gains; and resilience reserve $R(t)$ assessing cross-domain confirmatory breadth. These indicators serve as analytical categories for historical and philosophical analysis rather than quantitative metrics, providing structured tools for comparative assessment of epistemic system viability. The framework thus offers both philosophical coherence and practical diagnostic capacity without requiring the mechanistic claims of level 3.

Preserving the Insights: When we say "dispositions are compression algorithms" or "generative priors are Markov blankets," read these as capturing functional roles rather than ontological identities. A disposition functions like a compression algorithm: it reduces redundancy, encodes regularities, enables prediction. Whether it literally performs Huffman encoding or some neurally-implemented equivalent doesn't affect the philosophical point about what makes it successful or brittle.

Empirical Openness: This framework generates empirical predictions that could be tested. If consciousness really does track hierarchical compression processes, we should find neural signatures that correlate with compression improvements. If brittleness really does measure information leakage, we should be able to quantify it in failing belief systems. But even if specific empirical predictions fail, the conceptual framework for thinking about knowledge, truth, and existence retains value as a philosophical contribution.

Addressing the Framework-to-Payoff Concern: A legitimate worry about FEP-based frameworks concerns the ratio of elaborate formalism to novel predictions. Evolution's tautological core is tiny relative to its predictive machinery; the concern is whether this framework inverts that ratio. We acknowledge this challenge and offer three responses.

First, we identify a specific, surprising prediction: the coercion-before-collapse signature. The framework predicts that knowledge systems approaching brittleness thresholds should exhibit a characteristic temporal pattern: coercive overhead (resources devoted to suppressing disconfirming information) should rise

before model complexity explodes and *before* patch velocity accelerates. This is because coercion is the cheapest initial response to prediction error accumulation, becoming insufficient only as errors compound. This prediction is non-obvious: one might expect complexity increases or accelerating patches to come first. If historical analysis of collapsing paradigms (Ptolemaic astronomy, phlogiston chemistry, Lysenkoist biology) fails to show this coercion-first signature, the framework's brittleness dynamics would be disconfirmed.

Second, AI convergence provides a natural experiment. As artificial systems achieve genuine reasoning (not merely statistical pattern matching), the framework predicts they should converge on compression structures resembling formal logic and mathematics, not through training on human-generated data, but because these represent optimal compression of constraint relationships. If AI systems developed in isolation from human mathematical traditions achieve robust performance through radically different logical frameworks that don't reduce to or translate into human mathematics, this would constitute strong evidence against compression-driven convergence. The framework predicts substrate-independent convergence; radical divergence would disconfirm it.

Third, we distinguish convergence sources more carefully. Cross-cultural convergence in mathematics might reflect shared biology rather than compression optimality. But the framework makes a sharper prediction: structural learning should show asymmetric ease. Concepts that represent genuine compression joints (thermodynamic constraints, logical necessities, mathematical structures) should be easier to learn, more resistant to forgetting, and more reliably transmitted across generations than concepts of equal complexity that represent arbitrary conventions or false compressions. This predicts measurable cognitive asymmetries: learning that "fire produces heat" should show different retention curves than learning that "Thursday is named after Thor." If no such asymmetries exist, the distinction between structural and statistical patterns loses empirical grounding.

What Would Disconfirm the Optimal Constraint Configuration? We state explicitly: the Optimal Constraint Configuration concept would be disconfirmed by evidence that successful knowledge systems can achieve sustained low brittleness through compression strategies that do not converge. If independent civilizations (or AI systems, or future human cultures) developed equally robust, equally low-brittleness epistemic frameworks built on fundamentally incompatible logical or mathematical foundations, this would falsify the claim that compression optimality under shared constraints produces convergent structure. The Optimal Constraint Configuration is not merely "whatever works"; it is the claim that what works converges because optimal compression of shared constraint landscapes has unique solutions. Radical, stable pluralism in foundational frameworks would refute this.

With this methodological clarification in place, we can proceed to develop the framework confident that its philosophical insights don't collapse even if particular empirical implementations prove incorrect.

Having established the information-theoretic foundations of existence and the role of compression in naturalized epistemology, we now turn to the architectural structures that make bounded existence possible. Markov Blankets provide the formal mechanism by which systems achieve statistical separation from their environment while maintaining adaptive coupling, creating the conditions for computational closure and emergent autonomy.

3. Markov Blankets: The Architecture of Existence

3.1 What Is a Markov Blanket?

Consider a living cell. Its membrane separates "inside" (genes, metabolism) from "outside" (hostile chemistry). The membrane has sensors (receptors detecting nutrients) and actuators (secretions affecting environment). Crucially, the cell's internal processes depend only on what crosses the membrane, not directly on the external world. This is a Markov blanket: a statistical boundary creating conditional independence.

Formally, for a system with states partitioned into: - Internal states (μ): The "inside" of the entity - External states (η): The "outside" world - Sensory states (s): Detecting external changes - Active states (a): Affecting the external world

A Markov blanket exists when:

$$P(\mu \mid s, a, \eta) = P(\mu \mid s, a)$$

Internal states depend only on the blanket (sensory and active states), not directly on the external world. This creates conditional independence: the hallmark of autonomous existence.

This boundary formation is not merely a biological accident but a consequence of basic physical constraints. In any system where causes have effects, interactions propagate through neighbors, and states carry information about one another, statistical screening-off becomes inevitable. Certain configurations of matter will necessarily screen off interior from exterior states, such that the boundary becomes the sole causal mediator. Thus, the Markov blanket is not a label applied by an observer but a causal structure that naturally emerges from local interaction constraints.

3.2 Rainforest Realism: Blankets Delineate Real Patterns

Markov blankets are enacted, not discovered. They emerge when certain configurations of matter successfully maintain statistical boundaries against entropic dissolution. This perspective aligns with what Ladyman and Ross (2007) call *Rainforest Realism*: the ontology of the world is not a sparse desert of fundamental particles but a rich layering of real patterns at every scale where information can be compressed. A Markov blanket does not invent a useful fiction; it isolates a real pattern that carries genuine information load (what survives the compression process because it enables prediction).

3.2.1 Explicit Ontological Criterion: What Earns Existence?

This framework provides a clear answer to the fundamental metaphysical question: *What earns existence?* The answer is causal autonomy via computational closure.

A system exists as a distinct entity when it achieves a statistical boundary that makes its internal dynamics conditionally independent of its exterior. This is not merely about epistemic usefulness; it is about causal power. While Dennett (1991) rightly identified reliable compression as the hallmark of *real patterns*, we contend that nature itself enforces these compressions through thermodynamic constraints. *Causal Autonomy* occurs when a system's future states are fully

predictable and controllable, to a specified tolerance, using only variables defined at its own boundary.

3.2.2 Engagement with Mereology: Why Nihilism is Obsolete

This definition directly confronts mereological nihilism: the view that composite objects (like tables, cells, or hurricanes) do not exist, and that only fundamental particles arranged table-wise exist. The nihilist argues: *"When did this glass begin? No new matter was created, just rearranged soup. The 'glass' is just a mental label we project onto the atoms."*

Our response frames nihilism as obsolete rather than merely false. The "arrangement" of particles is not a neutral fact; specific arrangements create new causal constraints on information flow. When a boundary condition creates conditional independence (such that intervening on the boundary controls interior dynamics), the arrangement has achieved causal autonomy. To deny the existence of wholes at this point is to claim that causal efficacy is insufficient for existence, a standard that would render even fundamental particles (defined solely by their causal roles in field theory) non-existent.

Consider the internet. If we accept that only the micro-substrate exists, we are forced to conclude that the internet does not exist, only electrons in routers exist. Yet one cannot explain a server crash by analyzing electrons; one must analyze the network topology. The topology is a real causal agent. You can intervene on network-level variables (routing protocols, bandwidth allocation, firewall rules) to control outcomes in ways that electron-level interventions cannot achieve. To deny the existence of the whole is to deny the cause of the crash.

The "soup" is not made of self-subsistent things; it is made of constraints. Particles themselves are just nodes in relational structures (quantum fields). Thus, there is no "fundamental" level of particles that is more real than the "emergent" level of wholes. Both are *real patterns* that achieve computational closure at their respective scales. The universe is not flat; it is layered. And those layers are as real as the bedrock.

Examples Across Scales:

Scale	Entity	Markov Blanket	Internal States	Sensory/Active States
Molecular	Cell	Phospholipid membrane	Genes, metabolism, proteins	Ion channels, receptors, secretions
Neural	Brain region	Synaptic connections	Local processing circuits	Axonal inputs/outputs
Cognitive	Concept	Attentional filter	Compressed representation	Pattern recognition triggers, behavioral outputs
Social	Institution	Bureaucratic procedures	Internal decision-making	Public-facing policies, enforcement
Epistemic	Stable Concept	Definitional boundaries	Compressed causal model	Recognition criteria, licensed inferences

Recent neurobiological work identifies Markov blankets operating in canonical microcircuits and nested neural hierarchies (Hipólito et al. 2021), providing empirical grounding for the claim that blankets form across multiple organizational levels. These findings suggest the blanket architecture is not

merely a useful mathematical abstraction but reflects actual partitions in biological self-organizing systems.

The Radical Implication: What "exists" as a unified entity depends on which Markov blanket configuration you employ, though not all blanket configurations succeed. Cells exist for systems with cell-detecting blankets. Quarks exist for systems with quark-detecting blankets. Gods exist for systems with god-detecting blankets.

Defending Blanket-Relative Ontology:

This claim may seem to collapse into pure relativism or idealism, but it doesn't. The key is understanding that while blankets are enacted rather than discovered, not all enactments succeed. Reality imposes severe constraints on which blanket configurations achieve computational closure.

Consider three cases:

1. **Cells** (successful blanket): The phospholipid membrane creates genuine conditional independence. Internal metabolic dynamics can be predicted from membrane states alone, without tracking every external molecule. It serves as the paradigm case because the statistical boundary (Markov blanket) coincides perfectly with a physical boundary (lipid bilayer), making the conditional independence physically robust. The blanket achieves computational closure; it works. This is why cells persist across billions of years and countless environments.
2. **Phlogiston** (failed blanket): Attempts to draw a blanket around "phlogiston content" fail catastrophically. You cannot predict combustion outcomes using only phlogiston-level variables; oxygen levels, molecular structure, and thermodynamic conditions leak through constantly. Specifically, the weight gain of metals during calcination contradicted the prediction that phlogiston (matter) was released. Proponents attempted to save the theory by proposing that phlogiston possessed *negative weight*: a classic example of **Reactive Complexity** where auxiliary hypotheses ($M(t)$) proliferate to patch accurate predictions. However, this information leakage (the need to track external mass changes and invent ad-hoc physical properties) eventually shattered the blanket. The blanket never closes. This is why phlogiston was abandoned.
3. **Quarks** (successful but scale-dependent blanket): For particle physicists, quarks form a viable blanket; the Standard Model achieves computational closure at that scale. Quarks became "real" not by direct observation but because postulating them allowed the Standard Model to achieve computational closure, predicting particle interactions with high fidelity where previous models leaked information. For ecologists studying predator-prey dynamics, quarks are irrelevant; the blanket is drawn at the organism level. Both blankets work for their respective purposes.

The crucial insight: blanket-relativity is not antirealism. While the decision to draw a boundary is enacted by the observer or system, the viability of that boundary is determined by mind-independent constraints. We are free to draw a Markov blanket around "phlogiston," but we cannot force that blanket to achieve computational closure. The "terrain" of reality dictates which maps succeed.

Reality acts as the resistance to our attempted compressions. When we say a blanket is "enacted," we mean the attempt is internal to the system; when we say a blanket "exists," we mean the attempt has survived the filter of thermodynamic constraints. Phlogiston fails because the combustion process genuinely involves

oxygen, a fact about the world that generates high prediction error for any system attempting to ignore it. The ontology is relative to the blanket, but the success of the blanket is objective.

Ontological Pluralism with Objective Constraints: Different purposes require different blankets (quarks for physics, organisms for ecology, institutions for sociology), but within each domain, reality ruthlessly selects which blankets persist. This is neither naive realism (there is no single correct ontology) nor pure relativism (most attempted blankets fail). It is constrained pluralism; multiple viable ontologies exist, but viability is determined by reality's structure, not by our choices.

A methodological clarification: we distinguish between "Pearl blankets" (instrumental tools for statistical inference about systems) and "Friston blankets" (claims about the ontological boundaries of self-organizing systems) (Bruineberg et al. 2022). Our framework operates primarily at the functional level: Markov blankets characterize how systems achieve computational closure and maintain conditional independence, whether or not they correspond to specific physical boundaries. This functional interpretation avoids metaphysical overreach while preserving explanatory power.

The Biological-Epistemic Isomorphism: The parallel between biological and cultural-epistemic Markov blankets is not metaphorical but structural:

Biological Example	Cultural-Epistemic Example	Shared Mechanism
Cell membrane (phospholipid bilayer) blankets the interior from hostile chemistry outside; new causal level emerges (genes, metabolism, reproduction)	"…is an infectious disease" draws blanket around pathogen-host interactions, insulating public health reasoning from miasmas, humors, spirits; new causal level emerges (transmission chains, sterilization protocols, vaccines)	Both are coarse-grainings that minimize prediction error/free energy/ brittleness at the higher level
Cell receptor proteins = sensory states detecting nutrients/ threats	Recognition criteria = sensory states detecting instances ("has pathogen?")	Both detect relevant features across blanket boundary
Cell secretions/flagella = active states affecting environment	Licensed inferences/ interventions = active states affecting world	Both enable action based on internal model
Homeostasis = maintaining internal states despite external fluctuations	Functional entrenchment = maintaining predicate despite anomalies	Both resist dissolution through active maintenance

These analogies highlight functional isomorphisms rather than identical mechanisms. A cell membrane enforces thermodynamic boundaries through physical impermeability determined by molecular chemistry. Epistemic blankets like "…is an infectious disease" rely on pragmatic consensus and social coordination, permeable to negotiation and revision in ways biological membranes are not. The parallel lies in their shared function (both achieve computational closure by creating conditional independence), not in their implementation. This functional perspective prevents over-literal interpretation while preserving the genuine structural insight that both biological and epistemic systems solve the

same abstract problem: maintaining stable boundaries that enable higher-level causal dynamics to decouple from substrate details.

Once the blanket is in place, *you no longer reason from first principles every time*. Saying "COVID-19 is an infectious disease" instantly inherits isolation protocols, PCR testing, and ventilation engineering, just as a cell membrane instantly inherits billions of years of evolved receptor/secretion machinery. This is what a Markov blanket achieves: it lets the interior evolve under its own (much simpler) dynamics.

3.3 Computational Closure: When Emergence Succeeds

Definition: Computational Closure

A system achieves computational closure when coarse-grained macro-states form a complete, self-contained dynamical system. This requires three conditions:

1. **Lumpability:** Micro-states can be grouped into macro-states such that macro-dynamics are deterministic (same macro-state always transitions to the same next macro-state)
2. **Markovianness:** Future macro-states depend only on current macro-state, not on historical trajectory
3. **Causal shielding:** The macro-level is informationally closed from micro-implementation. Adding micro-level information does not improve macro-level prediction

When all three conditions hold, the macro-level constitutes an autonomous causal level. The ε-machine (using only macro-variables) and the υ-machine (with full micro-level access) achieve equivalent predictive accuracy.

This aligns with Dennett's definition of a real pattern as one that allows for descriptive efficiency better than a bit-map, but we add a crucial physical constraint: the universe itself must produce the boundary. The higher-level system becomes self-contained when you can predict future macro-states using only current macro-states, without tracking micro-details.

This hierarchical emergence arises through self-assembly. Kirchhoff et al. (2018) demonstrate that collectives of Markov blankets can self-assemble into global systems that themselves possess Markov blankets, creating nested boundaries from cells to organisms to social systems. Simulations demonstrate that hierarchical self-organization emerges naturally when microscopic elements have prior beliefs that they participate in macroscopic Markov blankets (Palacios et al. 2020). This suggests nested blanket hierarchies are not imposed from outside but arise spontaneously when components minimize free energy under appropriate constraints, providing a mechanistic account of how computational closure forms across levels.

Formalizing Causal Autonomy: Rosas et al. (2024) provide a rigorous framework for determining when emergence genuinely succeeds by comparing two optimal predictors. The ε-machine (epsilon-machine) uses only macro-level variables to predict future macro-states. The υ-machine (upsilon-machine) has full access to micro-level details and uses them to predict the same macro-states.

When these two predictors perform equally well (when ε-machine accuracy equals υ-machine accuracy), something remarkable has happened: the macro-level has achieved causal decoupling from the micro-level. The macro-variables contain all the information needed to predict their own future behavior. Substrate details have become causally irrelevant (though they remain constitutively necessary).

This is computational closure: the macro-level is "running code" rather than merely describing patterns in the substrate. This equivalence provides the rigorous criterion for a "Real Pattern" in Dennett's sense, transforming his "nontrivial compression" from a qualitative judgment into a quantifiable test. When adding micro-level details yields zero additional information about the macro-future, the macro-level description has captured genuine structure in reality rather than imposing convenient fiction. This equivalence admits degrees of robustness, formalized as lumpability. **Weak lumpability** holds when macro-dynamics work only for specific initial micro-state distributions: the compression succeeds in limited contexts. **Strong lumpability** holds when macro-dynamics work regardless of underlying micro-details: the compression achieves genuine substrate independence and persists across different physical realizations. Only strongly lumpable compressions qualify as objective features of reality; weakly lumpable ones are context-dependent approximations.

Examples:

Temperature (Successful Closure): - Micro: Positions and momenta of 10^{23} molecules - Macro: Single scalar (temperature) - The macro-variable (temperature) predicts thermodynamic behavior without tracking individual molecules - Causal closure: Heating water increases temperature increases pressure; the mechanism is shielded

Phlogiston (Failed Closure): - Attempted macro-variable: "Phlogiston content" - Failed lumpability: Cannot predict combustion outcomes without knowing oxygen levels, molecular structure, etc. - Information leaks through: Every new experiment reveals the blanket is porous - Brittleness accumulates: High $P(t)$ from constant patches

Connection to Generative Priors: "…is an infectious disease" achieves computational closure. Once you apply the predicate, you can reason about transmission, quarantine, sterilization (the higher-level causal dynamics) without tracking viral proteins, immune responses, etc. The predicate creates a new causal level.

3.4 ε-Machines: Optimal Blanket Constructors

ε-machines (epsilon-machines) are the mathematically optimal predictors: they compress past experience into the minimal set of causal states needed to predict the future.

Formal Definition: An ε-machine is a hidden Markov model that: 1. Maximally compresses the past (minimal number of states) 2. Maximally predicts the future (no lossless compression possible) 3. Achieves causal shielding (states are indistinguishable to the external observer)

Concrete Analogy: Chess Positions

Imagine two chess players trying to predict the outcome of a game:

- **The ∪-machine player** has access to the complete history: every move made, how long each player thought, what openings they studied, their heart rates, neuron firings in their brains. This player uses all micro-level information to predict the next move.
- **The ε-machine player** sees only the current board position: the macro-state. No history, no neural data, just the pieces and their locations.

When both players predict equally well, the board position has achieved computational closure. The macro-level (piece arrangement) contains all the information needed to predict future macro-states (subsequent positions). Crucially, once the current board position is known, the history of how the game arrived there is irrelevant for predicting legal moves. The game has "detached" from its substrate (the players' brains, their training, their moods) and runs purely on positional logic. This exemplifies causal shielding: the macro-state (board position) screens off historical details, achieving genuine computational closure. This effectively formalizes the *intentional stance* (Dennett 1987): treating the system as a macro-agent works not because it is a convenient fiction, but because the macro-dynamics have genuinely decoupled from the micro-details.

Rosas et al. (2024) formalize exactly this criterion: when your ε-machine (macro-only predictor) performs as well as the υ-machine (micro-informed predictor), you have discovered a level of organization that has achieved causal decoupling. The macro-level is "running code" rather than merely describing patterns in the substrate. This is the formal signature of successful emergence.

A different example illustrates continuous dynamical systems. Consider a traffic jam. A υ-machine attempting to predict when the jam clears by tracking every car's velocity and position faces an intractable computational task. An ε-machine tracking only traffic density and average flow rate achieves equivalent predictive accuracy with vastly fewer variables. When the ε-machine succeeds, the macro-variable "traffic density" has achieved computational closure. The micro-details of individual vehicles are constitutively necessary (no cars, no jam) but causally redundant; they have been screened off by the macro-level dynamics.

Why This Matters: - Dispositions are cognitive ε-machines: they compress experience into causal states (notions, beliefs) - Generative priors are cultural ε-machines: they compress collective experience into reusable causal tools - The optimal constraint configuration is the ultimate ε-machine: the minimal compression of reality's constraint structure

The Search Process: Organisms, communities, and species are ε-machine explorers, trying different compressions. The ones that achieve genuine computational closure while minimizing brittleness survive. This is not random but hill-climbing on the landscape of viable blanket configurations.

Quine's "web of belief" metaphor finds precise formalization in the ε-machine lattice structure described by Rosas et al. (2024). Each belief is a node in this lattice: a compression of experience with specific causal relationships to other beliefs. Central beliefs (logic, mathematics, basic thermodynamics) occupy positions of high structural compression: they encode fundamental constraints that many other beliefs depend upon. Revising them requires massive reorganization, changing prediction patterns throughout the entire system. Peripheral beliefs (today's weather forecast, a stranger's name) are shallow compressions with minimal dependencies. This explains Quine's observation that we revise peripheral beliefs readily when evidence conflicts, but resist changing core commitments. The resistance is not psychological stubbornness but information-theoretic necessity: central nodes bear higher revision costs. The web metaphor thus captures genuine computational structure: beliefs form a network where some compressions are load-bearing and others decorative.

3.5 Pragmatic Ontology: Same Information, Different Blankets

The Hot Dog Paradox: This illustrates how Markov blankets enact ontologies rather than discovering them.

The Information (Constant): - Bread-like substance encasing protein filling - Condiments, preparation method, consumption context - Physical/chemical properties unchanged

Different Markov Blankets (Variable):

Community	Blanket Boundary	Rationale	Ontological Commitment
Tax regulators	"Bread + filling = sandwich"	Consistent classification for revenue codes	Hot dog IS a sandwich
Culinary purists	"Requires two separate bread pieces"	Preserving fine-grained distinctions	Hot dog is NOT a sandwich
Structural engineers	"Continuous base with vertical walls"	Engineering load distribution	Context-dependent

Key Insight: Epistemic Equifinality. In Systems Theory, equifinality describes how different structural configurations can achieve the same steady state. The Hot Dog Paradox illustrates epistemic equifinality: different Markov Blankets (definitions) can achieve comparable levels of computational closure depending on the system's goal (taxation vs. cuisine vs. engineering). Each community draws the boundary where it reduces brittleness for their purposes. The information hasn't changed; the coarse-graining has. This is not arbitrary: each blanket faces pragmatic testing. However, it is pluralistic: multiple viable configurations exist.

Rosas et al. (2024) formalize this insight: valid coarse-grainings form a mathematical lattice structure, where multiple macro-level compressions can achieve strong lumpability for the same underlying system. The existence of this lattice explains epistemic equifinality: different purposes may select different points in the space of viable compressions, all of which genuinely achieve computational closure.

This lattice structure also provides the formal justification for Rainforest Realism. Because different physical substrates can realize computationally equivalent ε-machines, reality genuinely supports multiple valid ontologies enacted at different scales. This is a **Pragmatic Realism**: the *choice* of which blanket to draw is pragmatic (driven by the system's goals), but the *validity* of that blanket is realistic (constrained by strong lumpability). A "hot dog" is a valid object only if the hot dog variable predicts its own future better than noise. The rainforest is populated by all such valid closures.

Connection to Truth: The optimal constraint configuration doesn't dictate "the one true hot dog ontology" but rather the set of boundary-drawing strategies that achieve genuine computational closure with minimal brittleness. Different purposes require different closures. The Pluralist Frontier of the optimal constraint configuration is the zone where the constraint landscape is flat enough to support multiple, equally viable coarse-grainings (regions where equifinality holds).

Implication: Ontological disputes often aren't about facts but about which coarse-graining serves which purpose. The universe doesn't care if a hot dog is a sandwich, but food safety inspectors might need to draw that blanket for regulatory coherence.

3.6 From Notion to Generative Prior: The Blanket Formation Process

Terminological Note: We use "Generative Prior" throughout this paper to refer to successful, stabilized compressions that function as cognitive priors in predictive processing. Glenn (2025) uses "Standing Predicate" to refer to the same phenomenon. These terms are functionally identical: both denote beliefs that have achieved sufficient stability and compression efficiency to serve as foundational elements in a knowledge system, used to evaluate new claims rather than being evaluated themselves. We use "Generative Prior" here for its connection to Bayesian mechanics and information theory, while "Standing Predicate" emphasizes the linguistic and social dimensions. The choice is rhetorical rather than conceptual.

3.6.1 Notions as Proto-Markov Blankets

Before crystallizing explicitly, a belief exists as a tentative boundary-drawing attempt: what we call a notion. A notion is the brain's exploratory effort to see if a statistical boundary can be successfully maintained around certain patterns.

The Robinson Crusoe Insight: Generative priors don't require social coordination to form; they require only multimodal integration within a single agent navigating a constraint-rich environment.

3.6.2 Two Pathways to Valid Notions

The distinction between statistical and structural regularities (Section 2.5) explains why some notions can achieve validity from limited evidence while others require extensive testing:

Statistical Path (Requires Repetition): 1. **Multiple encounters:** Animal repeatedly shows certain behaviors 2. **Pattern extraction:** Brain notices correlation (this shape → barking sound) 3. **Gradual strengthening:** Each additional instance reinforces the compression 4. **Threshold crossing:** After sufficient trials, disposition stabilizes 5. **Example:** "Dogs are friendly" requires many positive encounters to overcome individual variation

Structural Path (Can Succeed from Singular Instances): 1. **Multimodal coherence detection:** Brain recognizes mutually constraining relationships 2. **Structural integration:** Components aren't just correlated but necessarily linked 3. **Immediate validity:** The pattern's internal coherence validates it without requiring repetition 4. **Example:** Touching fire once → permanent valid belief "fire burns"

3.6.3 Why Singular Instances Can Suffice

When Crusoe first encounters fire, he doesn't need hundreds of trials because fire exhibits structural coherence: - Combustion necessarily releases energy (thermodynamic constraint) - Energy necessarily manifests as heat/light (physical constraint) - Heat necessarily transfers to touching skin (causal constraint) - The process necessarily requires fuel (chemical constraint)

These constraints are mutually reinforcing. Recognizing any subset activates expectations about the others. This is why even a child touching a hot stove once

forms a lasting, valid compression: the pattern has structural integrity independent of frequency. Developmental psychology suggests that while statistical association often comes first, the "Aha!" moment of structural recognition represents a distinct cognitive phase shift where mutual constraints are detected.

Contrast: Why Statistics Sometimes Required:

Not all patterns have this structural character. "Ravens are black" is a statistical regularity without structural necessity; there's no thermodynamic or logical reason ravens couldn't be white. Such compressions require extensive sampling to distinguish genuine patterns from accidents of limited experience.

3.6.4 The Solitary Agent Formation Process

1. **Sensory encounter:** Pattern presents across multiple modalities
2. **Pattern type detection:** Brain assesses whether components show mere correlation (statistical, requires repetition) or mutual constraint (structural, can validate from limited data)
3. **Blanket formation attempt:** Brain draws tentative boundary around pattern
4. **Pragmatic testing:** Actions based on proto-blanket face reality
5. **Validation assessment:** Structural patterns strengthen immediately if internal coherence detected; statistical patterns strengthen gradually if predictions work across multiple trials
6. **Functional entrenchment:** Successful blanket becomes automatic disposition

Innovation Through Structural Recognition:

This explains how genuine innovations arise. A thinker detects structural coherence in a novel configuration before statistical validation.

Newton seeing a falling apple and recognizing universal gravitation didn't require thousands of falling objects. He recognized the structural relationship between terrestrial and celestial motion. The mathematical constraints (inverse-square law, conservation principles) exhibited internal coherence that could be validated through theoretical analysis before extensive empirical testing.

The Network Begins Inside: Even Crusoe alone has a "network": his visual, tactile, olfactory, and thermal subsystems must agree on "fire." The generative prior emerges when these internal streams achieve computational closure around a shared boundary. For structural patterns, this closure can succeed rapidly; for statistical patterns, it requires iterative refinement.

Social Transmission Accelerates Both Pathways: When language allows, successful blankets can be transmitted as linguistic handles: - "Fire burns" (structural) → transmitted with immediate credibility - "Mushrooms with red caps are poisonous" (statistical) → transmitted with caution, requires verification

But the fundamental process (drawing boundaries, testing them against reality, keeping those that close) operates identically in solitary and social contexts. The difference is whether the pattern's structure permits rapid validation or demands extensive statistical accumulation.

3.6.5 From Cognitive Mechanisms to Social Truth

A potential confusion must be addressed. Glenn (2025) distinguishes generative priors by their relationship to networks (Consensus Network versus Apex

Network). This paper distinguishes patterns by their cognitive processing requirements (statistical versus structural). Are these competing typologies?

No. They describe different dimensions of the same epistemic landscape. The two-way distinction (statistical versus structural) describes cognitive mechanisms: how the brain processes raw data to form beliefs. Network alignment describes validation states: how we determine whether those beliefs count as justified knowledge within a social context.

These dimensions interact but remain analytically distinct. Statistical and structural learning represent internal computational strategies. Consensus Network alignment represents external, intersubjective warrant. Understanding this relationship clarifies how individual cognition connects to collective epistemology.

Statistical Learning and Social Conformity:

Most cultural knowledge propagates through statistical learning. You hear "property rights are fundamental" or "democracy is best" thousands of times across family, education, media. Your brain detects frequency patterns, extracts correlations, builds compressions matching community norms. This achieves Consensus Network alignment automatically through repetition.

Phenomenologically, such knowledge feels natural and obvious. You "know" these truths without necessarily grasping their structural justification. This is System 1 processing (Kahneman 2011): unconscious, automatic, relying on signal redundancy. The knowledge rests on pattern matching rather than causal understanding.

The risk: if the Consensus Network encodes false compressions ("Miasma causes plague," "Phlogiston explains combustion"), statistical learning traps you in shared delusion. You align with the network by matching its patterns, regardless of whether those patterns correspond to reality's constraint structure. Statistical learning maintains network cohesion but provides no error-correction mechanism when the entire network drifts from truth.

Structural Learning and Objective Discovery:

Some knowledge arises through detecting structural coherence, potentially before or even against social consensus. Robinson Crusoe touches fire and immediately recognizes thermodynamic necessity linking combustion, heat, and pain. The pattern's internal coherence validates it without requiring social confirmation.

Phenomenologically, this produces the "Aha!" moment: "I see why this must be true." This is System 2 processing, conscious and attentional, recognizing necessity rather than mere correlation. Such learning can align with the Apex Network (objective reality) even when it conflicts with the Consensus Network.

This pathway enables breaking false consensus. When Einstein challenged Newtonian absolute simultaneity, he wasn't using statistical learning to absorb the dominant view. He detected a logical inconsistency in the relationship between light speed, simultaneity, and reference frames. His structural insight eventually pulled the Consensus Network toward the Apex Network. Without this mechanism, cultural knowledge systems would have no way to correct shared errors.

The Complementary Relationship:

Statistical learning maintains Consensus Network stability. It transmits accumulated knowledge efficiently across generations, preserves social cohesion,

and allows most people to function using compressions validated by their community without personally verifying each one.

Structural learning drives Consensus Network evolution toward the Apex Network. It provides the variation needed for cultural knowledge to improve. When individuals detect structural patterns that conflict with consensus, they generate proposals for network revision. Most such proposals fail, but successful ones constitute genuine epistemic progress.

Neither mechanism is universally superior. Statistical learning efficiently transmits validated knowledge but cannot detect when validation was mistaken. Structural learning can discover novel truths but often produces false insights that feel compelling yet prove brittle under testing. Both serve essential functions in collective knowledge production.

The framework thus synthesizes individual and social epistemology. The brain's two pattern-detection strategies map onto different network dynamics. Statistical learning explains how communities maintain stable shared knowledge. Structural learning explains how communities correct errors and discover new truths. Together, they account for both the persistence and the evolution of cultural knowledge systems.

Having established how Markov blankets enable computational closure and create autonomous causal levels, a question arises: which cognitive processes achieve conscious awareness? The statistical/structural distinction developed in Section 2.5 becomes crucial here. We propose that consciousness specifically tracks structural pattern recognition (the detection of bidirectional constraints) while statistical processing operates unconsciously.

4. From Information Processing to Consciousness

Having established how Markov blankets enable computational closure, we now turn to consciousness. The transition from mere information processing to subjective awareness remains the central puzzle of philosophy of mind. By applying our framework of structural vs. statistical pattern recognition, we can identify specific functional correlates of consciousness without claiming to fully solve the Hard Problem.

Scope of Claims: To be precise about our explanatory target, we distinguish three problems in consciousness research:

1. **The Hard Problem** (Chalmers 1996): Why does any information processing have phenomenal character at all? Why is there "something it is like" to see red rather than mere information processing without subjective experience?
2. **The Access Problem**: Which cognitive processes become available for conscious awareness, verbal report, and deliberate reasoning? Why can we introspect on some mental states but not others?
3. **The Functional Correlation Problem**: What functional or computational properties distinguish processes that reach consciousness from those that remain unconscious?

Our account addresses problems 2 and 3, not problem 1. We propose that structural pattern recognition (detecting mutual constraints between components)

is the functional process that correlates with access consciousness. Processes implementing statistical pattern matching, by contrast, operate automatically and transparently. This explains which information processing has phenomenal character, not why any processing has phenomenal character.

The following two sections address consciousness from complementary angles. Section 4A develops the theoretical framework, identifying functional correlates without claiming to solve the Hard Problem. Section 4B explores phenomenological applications and extensions of this framework. We identify structural pattern recognition and meta-blanket formation as functional correlates of consciousness without claiming to explain why these processes have phenomenological character. Phenomenology functions as the user interface for structural pattern recognition—the subjective signature of detecting mutual constraints between components. Whether this interface requires non-physical ontology, or whether the functional account exhausts consciousness, remains a separate metaphysical question beyond this framework's scope.

Our claim is functional and behavioral: systems achieving computational closure regarding their own structural inference processes will exhibit the reportable signatures of consciousness. This narrows the explanatory target without solving the Hard Problem. We are not asking "why does any information processing feel like something?" but rather "which information processing becomes accessible to awareness, and why?" This question admits functional answers grounded in computational differences between pattern recognition modes.

4A. Consciousness as Structural Pattern Recognition (Theory)

4A.1 The Core Distinction: Conscious vs. Unconscious Processing

Core Hypothesis: Consciousness may be the subjective experience associated with certain types of high-level information compression, particularly those involving structural pattern recognition happening in real-time, with particular salience for structurally coherent patterns over mere statistical regularities.

Central Distinction: Consciousness appears particularly engaged when the brain detects or attempts to detect structural coherence, when patterns present themselves as having internal necessity rather than mere correlation. Unconscious processing handles statistical pattern matching efficiently in the background, while conscious attention engages when structural relationships demand explicit reasoning.

Scope of Claim: We are not explaining why any information processing has phenomenal character (the hard problem). We are explaining which information processing correlates with reportable conscious access—the access problem. Our claim: structural pattern recognition is necessary for access consciousness. Processes detecting mutual constraints become available for verbal report and flexible behavioral control; statistical processing operates transparently. This generates testable predictions: disrupting structural inference circuits should selectively impair conscious access while leaving implicit pattern recognition intact.

4A.1.1 Negative Methodology for Consciousness Research

This framework builds knowledge from what we can confidently reject, complementing positive methodologies that attempt direct characterization of consciousness. While fallibilist, this approach provides robust constraints that prevent a collapse into relativism. By identifying functional processes that correlate with reportable awareness (structural pattern recognition, meta-blanket formation), we can reject theories that fail to account for these signatures while remaining agnostic about the underlying ontology of experience.

Two Modes of Pattern Processing:

Not all information processing reaches conscious awareness. We can distinguish between processes that operate primarily unconsciously and those that engage conscious attention:

Unconscious Processing (Statistical Pattern Matching): - Operates through frequency-based pattern recognition - Builds implicit compressions through repeated exposure - Examples: Walking, driving familiar routes, recognizing faces, grammatical intuitions - Phenomenology: Largely transparent to introspection ("I just know") - Information structure: Statistical correlations extracted from large samples - Requires: Multiple trials, gradual refinement, practice

Conscious Processing (Structural Pattern Recognition): - Engages when encountering patterns with internal constraint structure - Particularly active with novel patterns, contradictions, or structural relationships - Examples: Solving puzzles, understanding arguments, recognizing causal necessity - Phenomenology: Explicitly felt ("I see why," "this must be so," "something doesn't fit") - Information structure: Mutual constraints between components - Can operate: From limited data when structure is detected

The Conscious/Unconscious Boundary:

This distinction maps roughly onto the statistical/structural divide from Section 2.5. Consciousness appears particularly engaged when the brain detects or attempts to detect structural coherence (when patterns present themselves as having internal necessity rather than mere correlation).

Examples: - Learning to ride a bike (initially conscious structural analysis → becomes unconscious statistical refinement) - Recognizing a logical contradiction (conscious: structural incoherence demands attention) - Reading familiar words (unconscious: statistical pattern matching) - Understanding a proof (conscious: following chain of structural necessities)

Why This Matters for Phenomenology:

If consciousness tracks structural pattern recognition while unconscious processing handles statistical regularities, this explains several features of conscious experience:

1. **Novelty salience:** New structural patterns demand conscious attention (potential compression improvement)
2. **Contradiction salience:** Structural incoherence can't be ignored (breaks assumed constraints)
3. **Aha! moments:** Sudden recognition of structural relationships (see Section 4B.1 phenomenology table)
4. **Automation through practice:** Once structural understanding achieved, execution becomes statistical refinement (unconscious)

Connection to Active Inference Theories of Consciousness: Recent work grounds consciousness directly in Active Inference mechanisms. Laukkonen et al. (2025) propose that consciousness arises from a recursive self-evidencing loop in hierarchical systems. This requires three functional conditions: simulation of an epistemic field (world model), Bayesian binding (inferential competition to enter the world model), and epistemic depth (recurrent sharing of beliefs across levels).

Their framework aligns with our structural pattern recognition hypothesis. Consciousness may engage specifically when patterns exhibit sufficient internal constraint structure to support these recursive operations. Statistical regularities can be processed automatically without activating the full recursive loop. Structural coherence, however (precisely because components mutually constrain each other), demands the kind of inferential integration their theory describes.

The connection deepens when we consider Laukkonen et al.'s emphasis on counterfactual simulation and epistemic depth. Statistical learning creates flat predictions about the immediate sensory stream—what happens next given current input patterns. Structural learning, by contrast, enables detachment from the present to model "what if" scenarios: simulating alternative configurations, exploring constraint violations, testing structural relationships across hypothetical states. This capacity for counterfactual reasoning is precisely what Laukkonen identifies as creating "epistemic depth"—the recursive loop of simulating possible epistemic fields to select action policies. Our distinction between statistical (unconscious) and structural (conscious) processing thus maps onto their account: consciousness arises when the system engages in the counterfactual simulation that structural patterns both demand and enable.

Resonance with Classic Cognitive Science: This statistical/structural distinction also connects to foundational debates in cognitive science. Marcus (2001) famously argued that pure connectionist systems (neural networks relying on statistical pattern matching) struggle with the systematic, rule-based reasoning characteristic of human thought. While connectionist models excel at extracting correlations from data, they face difficulties with the compositional and algebraic operations that enable flexible generalization. Our framework provides an information-theoretic lens on this debate: what Marcus identified as the need for symbolic/algebraic operations corresponds to structural pattern recognition, detecting necessary relationships between mutually constraining components. Consciousness, on this account, may be the subjective signature of the system deploying these structural (algebraic) compressions, which are computationally distinct from the purely statistical regularities handled by unconscious subsystems. This suggests that the longstanding connectionist/symbolic divide reflects a genuine functional distinction in how patterns can be compressed, with consciousness tracking the structural mode.

Implications for Artificial Intelligence: This framework offers a diagnostic for the "hallucination" problem in current AI. Large Language Models operate primarily as ε-machines for statistical regularities, excelling at frequency-based pattern matching. While advanced Transformers can induce structural representations, their training objective (next-token prediction) prioritizes statistical likelihood over causal necessity. Hallucinations are often successful statistical mimicries that fail structural viability. A system predicting outcomes from corpus statistics rather than a causal world-model will inevitably "glitch" when statistical associations diverge from structural constraints.

4A.2 Hierarchical Compression and Meta-Awareness

Consciousness requires not just compression but meta-compression: compression of the compression process itself.

Three Levels: 1. **First-order processing:** Sensory data → compressed representations (largely unconscious) 2. **Second-order monitoring:** Awareness of dispositions (recognizing that you have a pattern-detector active) 3. **Third-order reflection:** Thinking about thinking (modeling your own modeling process)

Example: Recognizing Bias - First-order: "This person is untrustworthy" (disposition active) - Second-order: "I feel distrust toward this person" (aware of the disposition) - Third-order: "My distrust might be biased by their accent" (modeling the disposition's origins)

Only humans (as far as we know) achieve third-order regularly. This is meta-blanket formation: constructing a Markov blanket around your own Markov blankets, allowing self-modification.

Inner Screens and Imaginative Experience: Parr and Friston (2025) model this meta-level capacity through "inner screens"—internal boundaries with Markov blanket structure that function as classical information channels. These inner screens enable imaginative experience (planning, episodic memory, counterfactual reasoning) by allowing internally-generated content to employ the same spatial and conceptual reference frames used in ordinary perception. This provides a mechanistic account of how meta-awareness operates: the system constructs internal Markov blankets to model its own modeling processes, enabling the kind of third-order reflection that distinguishes human cognition.

The formation of meta-blankets is not merely metaphorical but physiologically grounded. Parr and Friston identify inner screens as functional Markov blankets within the cortical hierarchy that segregate imaginative planning from sensory perception. When you imagine biting into a lemon, you generate predictions about sourness without confusing them with actual taste input—the inner screen maintains the statistical boundary between internally-generated simulation and externally-driven sensation. This internal screening mechanism allows the system to manipulate its own dispositions as objects of thought: you can model your "distrust disposition" as a pattern to be examined, questioned, and potentially revised, rather than simply experiencing distrust as transparent reality. The meta-blanket thus enables the recursive operation where the system's predictive machinery becomes its own target of prediction, supporting the third-order reflection that characterizes metacognitive awareness.

The Self as a User Interface: Following Rosas et al. (2024), the "Self" is not a ghost in the machine but a control variable in the system's own high-level model: a user interface for the brain's self-regulation. It represents the brain's own lossy compression of its massive, distributed neural activity.

Just as a computer operating system represents billions of transistor states as a single "folder" icon, the brain compresses its complex somatic and cognitive states into a single variable: "I". This variable functions as a control parameter, allowing the system to predict and regulate its own future states without tracking every underlying neural process.

This variable is an ε-machine state: a simplified causal token that allows the system to predict its own future actions without tracking the firing of every individual neuron. The Self is not the neural hardware but the computational

software running on that hardware. It is the minimal effective theory required to predict the organism's future behavior.

This is not an illusion but a computational necessity. The brain must coarse-grain itself to operate at human-relevant timescales. The experience of being a unified "I" is what it feels like from inside this compression process, maintaining computational closure while billions of neural events churn beneath conscious awareness.

4A.3 Addressing (Not Solving) the Hard Problem

Traditional formulation: Why is there "something it is like" to process information? Why aren't we zombies? This is Chalmers' hard problem of consciousness: the explanatory gap between physical processes and subjective experience.

Correlates, not causes: A crucial clarification. We have identified a functional correlate of consciousness—structural pattern recognition appears to track the phenomenological boundary. But correlation is not causation. The explanatory challenge remains: why should detecting mutual constraints feel like *understanding* while detecting statistical correlations feels like nothing (or mere *familiarity*)? We have narrowed the space of functional properties to examine, but we have not explained why this particular functional property should generate subjective experience. The distinction between conscious and unconscious processing is functional, but the explanandum—why there is "something it is like" to perform one function but not the other—remains experiential and unresolved.

From the perspective of Ontic Structural Realism, this framing may itself be too concessive. If structure is ontologically fundamental—if there are no "things-in-themselves" beyond the relations they participate in—then a complete structural description of the compression process may exhaust the phenomenon rather than merely correlate with it. On this view, phenomenology is not a ghostly residue but the *informational view-from-within* of a Real Pattern maintaining its structural integrity. We remain agnostic on whether OSR fully dissolves the Hard Problem, but note that the "explanatory gap" may presuppose the very substance metaphysics that structural realism rejects.

Our Contribution (A Working Hypothesis):

We propose that consciousness relates to detecting and representing structural coherence rather than merely tracking statistical correlations. This isn't a solution to the hard problem, but it identifies a functional distinction that may map onto the phenomenological boundary between conscious and unconscious processing.

Recent work applies the Free Energy Principle directly to the hard problem, identifying affect (the feeling dimension of consciousness) as the subjective signature of free energy minimization—where decreases and increases in expected uncertainty are experienced as pleasure and unpleasure (Solms 2019). This suggests consciousness may track information-theoretic processes in a way that gives them phenomenological character, though why minimizing prediction error should feel like anything remains unexplained.

Alternative Formalization (Integrated Information Theory): A different approach starts from phenomenological axioms rather than functional principles. Oizumi et al. (2014) formalize Integrated Information Theory (IIT) by beginning with

intrinsic properties of experience (it is structured, integrated, definite) and deriving physical postulates about the mechanisms that could instantiate these properties. IIT identifies consciousness with the maximally irreducible conceptual structure generated by systems with high integrated information (Φ).

This framework can be contrasted productively with IIT as representing complementary methodological approaches. IIT adopts an intrinsic-nature-first methodology: it starts from phenomenological axioms about what consciousness is like (experience has definite boundaries, internal differentiation, integration) and works backward to identify the physical systems that could instantiate these properties. Our framework, by contrast, adopts a function-first methodology: it starts from the functional problem any bounded system faces (compressing reality to minimize prediction error) and works forward to identify the processes that would require conscious engagement (structural pattern recognition, meta-blanket formation).

These approaches converge from different directions. IIT focuses on irreducibility —consciousness arises where information integration creates a whole that cannot be reduced to independent parts. Our framework focuses on computational closure—consciousness arises where structural pattern recognition achieves causal autonomy from substrate. Both frameworks identify consciousness with causally powerful emergent structures operating at macro-levels. IIT measures this through Φ (integrated information); our framework measures it through successful compression achieving strong lumpability. The frameworks are complementary rather than competing: IIT addresses the intrinsic nature of experience, while our framework addresses the functional role consciousness plays in epistemic systems. A complete account may require both perspectives—understanding what consciousness is (IIT) and what it does (our framework).

The Distinction: - Unconscious processing: Statistical pattern matching (extracting correlations through frequency) - Conscious processing: Structural pattern recognition (detecting necessary relationships between mutually constraining components)

Why This Might Matter:

If phenomenology relates to structural pattern recognition, this would explain several otherwise puzzling features:

1. **Why singular experiences feel meaningful:** Structural patterns carry intrinsic relationships that don't require repetition to validate
2. **Why understanding feels different from familiarity:** Grasping structural necessity (conscious) vs. recognizing statistical pattern (unconscious)
3. **Why consciousness engages with novelty and contradiction:** Both demand structural analysis
4. **Why expertise becomes automatic:** Structural understanding converts to statistical refinement

The Phenomenological Texture of Structural Recognition:

Consider the qualitative difference between: - **Recognizing a face** (unconscious statistical matching: no sense of "why") - **Understanding why a proof works** (conscious structural analysis: sense of necessity)

The second has phenomenological character precisely because it involves representing constraint relationships: grasping that components must relate in certain ways. Perhaps phenomenology is what representing structural constraints feels like, while mere statistical correlation tracking proceeds unconsciously.

Consider the puzzle of mental representation: when one visualizes a triangle, a neuroscientist observing the brain sees only distributed firing patterns. The question "where is the triangle?" is a category error. The triangle exists at the level of the ε-machine—a coarse-grained macro-state that achieves computational closure over the neural substrate. The subjective experience of the triangle is the view from within that closure. It is "real" because it supports geometric inferences that the raw neural firing patterns cannot explicitly represent without the macro-level compression.

A functional clarification may help: phenomenology appears to function as the system's method of distinguishing between routine prediction errors (which can be smoothed out statistically over time through unconscious adjustment) and structural violations (which threaten the integrity of the model itself and demand immediate attention). Unconscious processing handles the former; phenomenology marks the latter. The subjective intensity of experiences like pain or epiphany may correspond to the magnitude of the free energy gradient associated with a structural constraint. While this does not explain the intrinsic nature of qualia (why the signal feels like *this*), it explains the functional necessity of intensity: the system requires a non-ignorable signal to disrupt automatic processing and force model reorganization. Phenomenology serves as the interface for high-stakes model revision.

What This Doesn't Explain:

We must be clear about the limits of this account:

- **Why structural detection feels like anything at all:** The hard problem remains (why any functional process has subjective character)
- **The specific quality of qualia:** Why red looks like *this* rather than something else
- **The unity of consciousness:** How distributed structural recognitions bind into unified experience
- **The possibility of zombies:** Whether systems implementing these functions necessarily have phenomenology

Our Position:

This framework identifies a functional distinction (statistical vs. structural pattern recognition) that appears to track the conscious/unconscious boundary. If consciousness really does engage with structural coherence specifically, this narrows the explanatory target.

We're not asking "why does any information processing feel like something?" but rather "why does detecting mutual constraints between components feel like something?"

That's still a hard problem, but it's a more precise one. And it suggests consciousness isn't an arbitrary add-on to cognition but tracks a genuine functional distinction in how patterns can be recognized.

Not Eliminativism, Not Mysterianism:

We're not denying consciousness exists (eliminativism). We're not claiming it's forever inexplicable (mysterianism). We're offering a naturalistic framework that respects both the reality of phenomenology and the difficulty of explaining it.

Consciousness may be the subjective signature of structural pattern recognition: what mutual constraint detection feels like from inside the system performing it.

Why detecting structural constraints should feel like anything remains an open question this framework doesn't fully answer. We've identified a relevant functional distinction; we haven't solved the hard problem.

4B. Applications and Extensions

4B.1 Phenomenology of Compression

The distinction between statistical and structural processing offers a detailed map of phenomenological states. We propose that specific qualitative experiences correspond to distinct states of the compression engine:

Conscious Experience	Information-Theoretic Process	Pattern Type
"Aha!" moment	Sudden recognition of structural coherence: components snap into mutually constraining relationship	Structural
Confusion	Inability to find structural pattern: high prediction error persists without coherent explanation	Mixed
Understanding	Achieving structural compression: grasping why components must relate as they do	Structural
Certainty	Structural necessity recognized: pattern exhibits internal constraint	Structural
Doubt	Statistical evidence conflicts or structural coherence unclear	Mixed
Boredom	Maximal compression achieved: no new structural insights available (perfectly predictable)	Both
Curiosity	Hints of structural pattern not yet grasped: optimal zone for compression improvement	Structural
Flow state	Structural understanding guides action while statistical refinement operates unconsciously	Both
Anxiety	High prediction error without structural explanation: cannot find pattern that makes sense	Mixed
Familiarity	Statistical pattern matching sufficient: no structural analysis needed	Statistical

4B.2 Emotions as Free-Energy Gradients

Emotions are not irrational disruptions but dashboard readings of the epistemic engine's state:

Negative Emotions (High Free Energy): - **Anxiety:** Persistent prediction error without identified cause - **Frustration:** Model predicts actions should work, but

outcomes consistently differ - **Shame/Guilt:** Social model predicts acceptance, but feedback signals rejection - **Confusion:** Cannot compress incoming information into existing categories

Positive Emotions (Low Free Energy): - **Joy:** Predictions confirmed, model vindicated - **Relief:** Expected high free energy avoided - **Pride:** Social model predicts acceptance, feedback confirms - **Satisfaction:** Goals achieved as predicted

Motivational Emotions (Free Energy Gradients): - **Curiosity:** Moderate prediction error signaling compressible patterns (explore!) - **Fear:** High prediction error signaling danger (freeze/flee!) - **Anger:** Obstacle blocking expected state (remove/attack!)

From this perspective, emotions are not bugs but features: they make the costs of misalignment consciously accessible, motivating revision toward lower-brittleness configurations.

4B.3 Agency and Variation

Determinism Concern: If beliefs are compressions shaped by information, where is agency?

Resolution: Agency is not freedom from causation but internally-generated variation within constraint space. The variation originates from the system's own dynamics rather than external perturbation, though it remains causally constrained by the system's structure and history.

Three Senses of Agency:

1. **Metabolic Agency (Cells):** Active inference (changing environment to match predictions)
   - The cell extends pseudopods toward nutrients (active state)
   - This reduces prediction error (sensory state matches "food here" prediction)
2. **Behavioral Agency (Animals):** Exploration of action space
   - Try different behaviors, keep those that reduce free energy
   - Learning is ε-machine construction through trial-and-error
3. **Cognitive Agency (Humans):** Mental simulation and deliberate blanket modification
   - Imagine counterfactuals ("what if I revise this belief?")
   - Deliberately test alternative compressions
   - Meta-blanket agency: Modify your own information-processing architecture

The Mechanism of Cognitive Variation: Cognitive agency is possible because we compress the world into Standing Predicates that encode structural relationships (Section 2.5). Unlike purely statistical regularities that must be learned through repeated exposure, structural patterns capture causal dependencies that can be manipulated offline. When you imagine "what if disease is caused by invisible organisms?", you're not randomly mutating beliefs but systematically exploring how rearranging causal components would change predictions. This capacity for counterfactual simulation on structural constraints distinguishes cognitive agency from metabolic or behavioral exploration—you can test compressions in simulation space before bearing the thermodynamic costs of testing them in reality. The variation is not arbitrary but constrained by which reconfigurations preserve internal coherence while potentially improving alignment with reality's structure.

Free Will Recovered: The capacity to generate novel compressions (new functional propositions, heresies) before reality tests them. Humans can: - Propose new blanket configurations ("what if disease is caused by invisible organisms?") - Mentally simulate outcomes - Deliberately adopt high-cost positions to test them - Bear brittleness costs to explore the compression landscape

The Variation Engine: Individuals are variation generators; reality is the selection mechanism. The optimal constraint configuration is discovered through distributed exploration, not centrally imposed.

Important: This isn't libertarian free will (causally uncaused choices) but compatibilist agency (internally-generated variation within causal constraints).

Having explored how consciousness emerges from structural pattern recognition and enables agency through variation generation, we now examine the most stable compression structures that emerge from this process: logic and mathematics. These formal systems represent the deepest structural regularities discovered by distributed intelligence across generations, achieving maximal computational closure and serving as the backbone of the Apex Network.

5. Logic and Mathematics as Necessary Compression Structures

5.1 Why Logic Occupies the Core

Traditional view: Logic is a priori, transcendentally necessary, or conventionally chosen.

Information-theoretic view: Logic is the minimal compression structure required for any system capable of error-correction.

The Transcendental Argument:

Any system capable of error-correction must be able to distinguish success from failure. This requires recognizing when A and not-A cannot both be true (non-contradiction).

Chains of inference require transitivity: if believing A leads you to believe B, and believing B leads you to believe C, then believing A should lead you to believe C. Otherwise, your compressions fragment into isolated islands.

Together, these minimal requirements (non-contradiction and transitivity) form the core of classical logic. Logic is not metaphysically necessary in some Platonic sense. It is functionally prerequisite.

Any system that learns (compresses experience, updates on prediction error) must implement logical structure. Basic logical principles appear functionally necessary for any system that maintains coherent compressions. This is not a claim about metaphysical necessity but about functional prerequisites given the constraints of bounded information processing.

This requirement has a precise information-theoretic interpretation: logic ensures that operations at different scales remain consistent. If you perform a computation at the micro-level and then compress the result, you should get the same answer as if you compressed first and then computed at the macro-level. When this consistency fails, the macro-level description no longer tracks the

underlying dynamics, and the compression breaks down. Logical laws (like transitivity and non-contradiction) are thus the structural requirements for valid coarse-graining. They are not Platonic impositions but the conditions under which any "software layer" can reliably run on its physical "hardware."

Information-Theoretic Grounding: - Non-contradiction: Same input cannot compress to contradictory outputs - Excluded middle: Compression requires binary decision boundaries - Modus ponens: Compression chains propagate information - Identity: Compression requires stable reference

Revising Logic: Would require dismantling the error-correction mechanism itself. This generates infinite brittleness: the system would have no way to evaluate whether the revision succeeded or failed.

5.2 Mathematics as Optimal Compression

We propose that mathematical structures represent optimal compressions of structural regularities.

Examples:

π (Pi): - Compresses infinite information (circle's circumference/diameter ratio) - Into finite symbol with infinite precision - Necessarily determined by Euclidean geometry's constraint structure - Discovered independently across cultures (Babylonians, Greeks, Indians) because constraint structure is objective

Prime Numbers: - Compress information about multiplicative structure - Their distribution compresses deep regularities in arithmetic - Riemann Hypothesis (if true) would be ultimate compression of prime distribution

Group Theory: - Compresses symmetries across domains (crystals, particles, equations) - One framework compresses structure in chemistry, physics, music, cryptography - Unreasonable effectiveness because it captures genuine compression joints

Connection to Optimal Structure: Mathematics is part of the optimal constraint configuration: the maximally compressed representation of structural constraints that any sufficiently thorough compression must discover.

Philosophical Positioning: This view differs from Mathematical Platonism in key respects. Platonism holds that mathematical objects exist independently in an abstract realm, which our minds somehow access. We claim instead that mathematical structures are optimal compression protocols for describing constraint relationships; they are discovered rather than invented, but what is being discovered is the structure of the constraint space itself, not objects in a separate ontological realm. This aligns more closely with ontic structural realism (Ladyman and Ross 2007), which holds that structure rather than objects is ontologically fundamental. The "necessity" of mathematics derives not from inhabiting a timeless Platonic heaven but from the fact that certain compression strategies are uniquely optimal given the axioms and constraints. Independent discovery of π across cultures evidences objective structure, but what makes π necessary is the relationship it compresses (circumference to diameter in Euclidean space), not its existence as an abstract entity. This framework preserves mathematical objectivity without requiring a separate realm of mathematical objects.

Mathematical Fallibility: Even mathematics has its Negative Canon. Naive Set Theory, which allowed sets to contain themselves, generated Russell's Paradox, a

contradiction revealing that the blanket had leaked. The mathematical community didn't simply choose to revise set theory; they were forced to by the internal incoherence the system produced. Failed mathematical definitions, like failed scientific theories, represent compressions that could not achieve computational closure. They generated contradictions (information leakage at the logical level) and were selected against, demonstrating that mathematical knowledge, like empirical knowledge, is subject to the same selection pressures we describe throughout this framework.

5.3 The Unreasonable Effectiveness of Mathematics

Wigner's Puzzle: Physicist Eugene Wigner famously asked why mathematical structures discovered purely abstractly apply to physical reality with uncanny accuracy. Why should group theory, developed to study symmetries in abstract algebra, perfectly describe particle physics?

Information-Theoretic Answer: Mathematics and physics are exploring the same compression landscape from different angles: - Physics: Compress experimental observations - Mathematics: Compress structural necessities - They converge because both face the same constraint structure

Example: General Relativity - Einstein: "Find simplest equations describing gravity" - Mathematicians: "What's the geometry of curved spaces?" - Same compression achieved from different starting points - Convergence reveals objective structure (the Apex Network of geometry)

These logical and mathematical structures provide the rules for compression. We now examine how successful compression creates new causal levels through computational closure, demonstrating how macro-level causal autonomy emerges from micro-level interactions when information leakage is minimized.

6. Emergence Through Computational Closure

Classical mystery: How do qualitatively new properties (liquidity, life, consciousness) emerge from mere rearrangement of parts?

The answer follows from Section 3.3: emergence succeeds when a Markov blanket configuration achieves computational closure (lumpability, Markovianness, and causal shielding). When these conditions hold, the emergent level is as causally real as the base level.

Rosas et al. (2024) formalize this as causal decoupling: when the macro-level ε-machine achieves equivalent predictive accuracy to the micro-informed υ-machine, the macro-level has achieved genuine autonomy. This is strong emergence: not merely useful description but substrate-independent causal dynamics.

Failed emergence equals information leakage. Phlogiston's attempted macro-variable ("phlogiston content") failed lumpability: predicting combustion required oxygen levels, molecular structure, and other substrate details the macro-variable couldn't capture. The blanket was porous; brittleness accumulated until abandonment.

This grounds the transition to Section 7: the Optimal Constraint Configuration represents the set of compressions that achieve maximally robust computational closure across domains.

7. The Optimal Constraint Configuration as Ultimate ε-Machine

7.1 Synthesizing Information, Compression, and Truth

The optimal constraint configuration is the complete set of generative prior configurations that achieve minimum systemic brittleness: the intersection of all maximally viable compression structures. In information-theoretic terms, it represents the "ultimate" ε-machine (using the term to denote the limit of maximal compression rather than a final, static state). It functions as the thermodynamic attractor where information leakage is theoretically minimized.

Rosas et al. (2024) demonstrate that all valid coarse-grainings of a system form a mathematical **lattice**: a hierarchical structure of nested compression levels, where each node represents a different way to group micro-states into macro-states. Not all coarse-grainings are equally robust: some achieve only weak lumpability (working only for specific initial conditions), while others achieve strong lumpability (preserving macro-dynamics regardless of substrate details). The optimal constraint configuration corresponds to the optimal path through this lattice: the set of strongly lumpable coarse-grainings that maximize causal autonomy while minimizing computational complexity. Reality allows many valid maps (the full lattice), but the optimal constraint configuration represents those compressions that achieve genuine substrate independence. This is why objective truth is not correspondence to a single privileged description but the achievement of strong lumpability: the predicate holds regardless of the underlying micro-state distribution, making it a genuine feature of reality rather than an artifact of our perspective.

Terminological Note: While the mathematical structure is technically a lattice (a partially ordered set with strict hierarchical derivation), we use "optimal constraint configuration" to emphasize the socially shared and collectively discovered nature of these optimal compressions. This configuration emerges through distributed exploration by communities of knowers rather than individual deduction.

Ontological Status (Structural Emergent, Not Metaphysical Blueprint):

The optimal constraint configuration is not a pre-existing Platonic form or cosmic blueprint awaiting discovery. It is the pattern that necessarily crystallizes from the interaction between information-processing systems and environmental constraints. It is determined by constraints rather than our beliefs about them. Its existence is the existence of a determined pattern, not a transcendent entity.

The optimal constraint configuration is constrained by the Principle of Naturalistic Closure: a generative prior cannot achieve optimal status through mere internal coherence (Level 3 truth) but must unify special-science hypotheses without contradicting fundamental physics. This is why the optimal constraint configuration is not merely "what works for us" (pragmatism) or "what our community agrees upon" (social constructivism), but the set of real patterns that

achieve genuine integration across the hierarchy of sciences—from thermodynamics to biology to psychology. The constraint is not just thermodynamic viability but consistency with the mathematical structure of our universe.

Modal Determinacy: Given our universe's actual constraint structure (thermodynamics, logical consistency, biological requirements), the optimal constraint configuration appears to be the necessary optimal configuration, modally necessary relative to these constraints, though this claim remains theoretical and requires further empirical validation. However, in a universe with different fundamental physics or logical laws, a different optimal configuration would emerge. There is no super-cosmic structure independent of physical reality itself.

Analogy to Mathematical Necessity: Just as π is not "somewhere" waiting to be found but is a necessary consequence of Euclidean geometry's constraint structure, the optimal constraint configuration is a necessary consequence of reality's pragmatic constraint structure. Ancient Babylonians, Greeks, and Indians discovered π independently not through cultural transmission but because geometric constraints determine its value. Similarly, independent cultures converge on similar low-brittleness principles (reciprocity norms, property conventions, harm prohibitions) because these are structurally necessary for viable coordination, determined by objective pragmatic constraints.

The constraints exist first; the optimal structure they determine is a necessary implication. Historical filtering is how we discover this structure, not how we create it.

Collective Intelligence and Emergent Knowledge: Recent work on multi-agent active inference provides a concrete mechanism for understanding how the optimal constraint configuration emerges through distributed exploration rather than centralized design. Friston et al. (2025) demonstrate that collectives can be treated as emergent agents with their own Markov blankets, possessing synergistic information that no individual member encodes. In their model of flocking birds, the collective "knows" the predator's location and optimal escape direction through the coordination of individual responses, even though no single bird detects or represents this information. The flock exhibits genuine collective knowledge that emerges from distributed processing.

This constraint-determined structure aligns with what Dittrich and Kinne (2024) call "reality-filtered data." Information produced by surviving agents already embeds causal structure because only causally accurate compressions enable the agents generating that data to persist. The optimal constraint configuration can thus be understood as the attractor state in the space of possible compressions where information-theoretic efficiency is maximized. It represents the set of generative priors that constitute sustainable, generative models (in CEP's terms), discovered through distributed exploration of reality's constraint structure. This convergence mechanism explains how cultural and historical processes can approach objective truth without requiring direct access to a Platonic realm: agents exploring reality under survival pressure necessarily generate data reflecting causal constraints, and compressions of that data necessarily track those constraints.

This formalizes how the optimal constraint configuration operates. Just as a flock possesses information about environmental gradients that individual birds do not, a culture's network of generative priors can encode compressions that no individual fully grasps. The optimal constraint configuration represents the

collective computational closure achieved when communities of knowers, each exploring different regions of the compression landscape, generate a distributed set of predicates whose interactions reveal constraint structures invisible to isolated inquiry. It is neither a pre-existing blueprint awaiting discovery nor an arbitrary social construction, but the emergent pattern of collective intelligence navigating reality's constraints.

7.1.1 Logical Depth as Epistemic Capital

We must distinguish between mere compressibility (Kolmogorov complexity) and the value of the compression. A random string is incompressible; a string of a million zeros is highly compressible but trivial. The optimal constraint configuration consists of compressions that exhibit high *logical depth* (Bennett 1988). Logical depth measures the computational work (time and energy) required to generate a pattern from its minimal description.

Stable concepts (acting as shared generative priors) function as batteries of stored computational work. The predicate "$F = ma$" is brief, but its Logical Depth is immense: it encodes the integrated output of centuries of astronomical observation and calculus development. When an agent adopts a high-depth predicate from the consensus network, they inherit the result of this thermodynamic expenditure without performing the computation themselves. This formalizes the function of a generative prior: it encapsulates the computational history required to derive it. Truth, in this framework, is not merely efficient compression; it is the maximization of Logical Depth: the ratio of explanatory power to the thermodynamic cost of derivation. This explains why the optimal constraint configuration is an attractor: it represents the path of least resistance for minimizing future metabolic costs via the inheritance of past computational work.

Maximum Computational Closure as Thermodynamic Minimum:

In information-theoretic terms, the optimal constraint configuration represents the thermodynamic attractor in the phase space of possible compression systems: the configuration where information leakage is theoretically minimized. This is the limit state where Markov blankets achieve maximum alignment with environmental causal structure:

The Limit State: Maximum computational closure occurs when the internal model predicts the external environment with such accuracy that information leakage approaches zero. Not because the Markov blanket boundary vanishes, but because the compression achieves such high fidelity that the enacted boundary perfectly tracks genuine causal joints in reality.

Analogy: A perfect mirror doesn't eliminate the boundary between object and reflection, but the information crossing that boundary is transmitted with such fidelity that operationally, the distinction disappears. Similarly, the optimal constraint configuration is where conceptual boundaries (Markov blankets) achieve such high-fidelity compression that they mirror reality's constraint structure exactly.

Plateau, Not Necessarily Single Peak:

The optimal constraint configuration should not be understood as a single, final theory of everything. Rather, it is the complete set of maximally viable configurations: a high-altitude plateau on the fitness landscape. While some domains may have single sharp peaks (basic thermodynamics, core logic), others may permit constrained pluralism of equally low-brittleness systems (hot dog

taxonomy, aesthetic frameworks). Convergence is away from vast valleys of failure (the Negative Canon) and toward this resilient plateau of viable solutions.

Consider how this plateau manifests across different domains. In ethics, reciprocity norms exhibit striking cross-cultural convergence: the Golden Rule and its variants appear independently across civilizations because coordination constraints impose pragmatic necessities on any viable social system. Yet this convergence on core principles permits local variation in implementation details. Similarly, in aesthetics, tonal and atonal music systems represent different organizational frameworks that can achieve comparable coherence and expressive power within their respective constraint structures, neither uniquely privileged by reality's constraint landscape. In macro-epistemology, different property rights configurations—communal ownership systems versus individual title regimes—can minimize systemic brittleness under different ecological and technological conditions, achieving equifinality through distinct institutional architectures. The plateau thus captures both the objectivity of viability constraints (vast regions of configuration space generate catastrophic brittleness) and the pluralism of viable solutions (multiple paths can satisfy those constraints equally well). What matters is that systems occupy the high-altitude terrain where brittleness is minimized, not which specific coordinates they inhabit on that plateau.

7.1.2 What the Optimal Constraint Configuration Is and Is Not

To prevent misunderstanding, we distinguish the OCC from several alternatives:

The OCC is not a single final theory. Multiple compression strategies may achieve equally minimal brittleness for different purposes (the hot dog taxonomy example in Section 3.5 illustrates this). The OCC is the envelope of all such viable strategies: the boundary separating sustainable compressions from catastrophically brittle ones. It represents the constraint landscape's viable region, not a unique point within it.

The OCC is not observer-independent in configuration but is observer-independent in constraint. Which specific compressions a system adopts depends on its purposes (particle physics versus ecology, tax codes versus culinary traditions). But which compressions are viable at all is determined by reality's constraint structure, not by choice or convention. We are free to attempt any compression we wish; reality determines which attempts achieve computational closure.

The OCC is defined negatively rather than positively. We cannot specify in advance what the OCC contains, as this would require complete knowledge of reality's constraint structure. Instead, we discover it by eliminating what it excludes: the Negative Canon of compressions that generate unsustainable brittleness. The OCC is the residual structure that survives this elimination, revealed through historical filtering rather than deductive proof.

Reconciling "Plateau" and "Intersection" Characterizations: These descriptions are compatible when properly understood. The OCC is an intersection in the sense that it represents constraints shared across all viable systems: no viable system can violate thermodynamic laws, logical consistency, or basic coordination requirements without accumulating catastrophic brittleness. The intersection defines the boundary between viable and nonviable compression spaces. The OCC is a plateau in the sense that within these shared constraints, multiple specific configurations may achieve equally minimal brittleness. The intersection defines the plateau's boundaries; the plateau's surface represents the space of viable variation within those boundaries. The formal characterization $C_{opt} = \cap$

$\{W_k \mid V(W_k) = 1\}$ captures the intersection of all maximally viable configurations, which collectively form the high-altitude plateau where brittleness is minimized.

Addressing the Circularity Objection: A circularity worry arises: if truth is alignment with viable compressions, and viability is measured by persistence, haven't we defined truth as "whatever survives"? This would make truth hostage to power, historical accident, or mere stubbornness.

The response distinguishes contingent persistence from structural viability. A compression can persist contingently through resource extraction, coercive suppression of alternatives, or isolation from disconfirming evidence. But these strategies generate characteristic brittleness signatures (high $C(t)$, accelerating $P(t)$, rising $M(t)$) that predict eventual failure. Structural viability means persistence under open epistemic conditions—where error signals are not suppressed and alternatives can compete freely. The OCC is defined not by what happens to persist under arbitrary conditions, but by what would persist under idealized conditions of free inquiry and pragmatic testing.

This parallels how evolutionary fitness is not "whatever survives" but "whatever would survive under selection pressure in a given environment." A species artificially protected from predators may persist without being fit. A belief system maintained through censorship may persist without being true. The OCC represents the attractor under selection, not the current census of survivors. The circularity dissolves when we recognize that viability is defined by a counterfactual standard (persistence under open conditions) rather than actual historical persistence.

Ontologically Real, Epistemically Regulative:

A crucial distinction: The optimal constraint configuration is ontologically real: the objective, mind-independent structure of viability that exists whether we correctly perceive it or not, determined by constraints rather than our beliefs. However, epistemically it remains a regulative ideal. We can never achieve final confirmation that our consensus models perfectly map it; our knowledge is necessarily incomplete and fallible.

This dual status grounds a form of structural realism (there is an objective constraint structure) while preserving epistemic humility (we cannot claim certainty about fully capturing it). We propose this as a theoretical model for how objective knowledge is possible in a physical universe, acknowledging that the extent to which human knowledge systems actually achieve this alignment remains an open empirical question.

The optimal constraint configuration exists as π exists: determined by constraints, counterfactually stable across possible histories, discoverable through systematic exploration. But unlike a Platonic form, it is an immanent pattern. It is the negative space revealed when systematic pragmatic filtering eliminates unviable alternatives.

Formal Characterization (With Appropriate Caveats):

We can characterize the optimal constraint configuration as the intersection of all maximally viable world-systems:

$$C_{opt} = \cap \{W_k \mid V(W_k) = 1\}$$

Where C_{opt} = Optimal Constraint Configuration, W_k = possible configurations of generative priors, $V(W_k)$ = viability function (inversely related to brittleness metrics), and $\cap$ = intersection (common structure across all viable systems).

This formalism captures the concept but should not be mistaken for literal metaphysics. It represents the structural pattern that emerges from constraint-driven selection, not a pre-temporal mathematical object.

Terminological Note: Glenn (2025) uses "Apex Network" to refer to this same structure. The concepts are identical: both denote the objective, constraint-determined configuration of minimum systemic brittleness toward which viable knowledge systems must converge. We use "Optimal Constraint Configuration" here for its more technical, information-theoretic flavor, while "Apex Network" serves the same function with more natural phrasing. The choice is rhetorical rather than conceptual. Both terms describe the thermodynamic attractor in the phase space of possible belief systems, the intersection of all maximally viable configurations, existing whether discovered or not because it is determined by reality's constraint structure rather than by our beliefs about it.

7.2 Truth as Successful Computational Closure

Redefining Truth: A proposition is true (Level 1) if its predicates are part of the optimal constraint configuration: the optimal computational closure configuration. In Rosas et al.'s (2024) terms, objective truth corresponds to *strong lumpability*: the predicate holds regardless of underlying substrate or initial micro-state distribution. A weakly lumpable predicate works only for specific conditions: it may be locally useful but not objectively true. A strongly lumpable predicate works across all valid realizations: it has achieved genuine substrate independence and thus qualifies as objective truth. Truth is not arbitrary social construction but achievement of maximal causal autonomy in the compression lattice. The optimal constraint configuration is modally necessary relative to our universe's constraint structure, though alternative universes with different constraints would yield different optimal configurations. This counterfactual stability does not imply metaphysical necessity, but structural necessity given pragmatic constraints.

Epistemic Qualification: This framework provides a naturalistic account of objectivity, though our access to the optimal constraint configuration remains fallible and requires empirical triangulation. While the optimal constraint configuration exists as a constraint-determined structure (the set of compressions that minimize information leakage given actual pragmatic constraints), our knowledge of it is always mediated by historical filtering and subject to revision. Strong lumpability functions as a regulative ideal—the standard against which we assess compressions—but actual knowledge claims remain provisional. We can have justified confidence that certain predicates approach strong lumpability (those demonstrating low brittleness across diverse contexts and extended time periods), but we cannot claim infallible access to the complete optimal constraint configuration. The framework thus preserves robust realism about truth's objectivity while maintaining appropriate epistemological humility about our current knowledge.

Three Levels Revisited Through Information Theory:

Level	Information-Theoretic Characterization	Phenomenology
Level 0 (Contextual Belief)	Unjustified, unfalsifiable belief lacking compression structure	Feels true due to social or emotional factors
Level 3 (Coherence)	Internal consistency within a local compression scheme	Feels true within the system
Level 2 (Justified)	Compression validated by low brittleness in practice	Rationally confident it's true
Level 1 (Objective)	Part of the optimal constraint configuration: optimal compression of constraint structure	Would be true even if we never discovered it

Example: Heliocentrism - Level 3: Coherent within Copernican framework (even before validation) - Level 2: Justified once observations confirmed lower brittleness than geocentrism - Level 1: Objectively true because it's part of optimal compression of gravitational constraints

Example: AI Development Current trends suggest potential brittleness in AI systems, though this requires further empirical validation. Large language models demonstrate impressive pattern recognition but may exhibit brittleness in novel domains where their statistical training data provides insufficient coverage.

Quine argued that truth is immanent to our best theory—determined by what belief revisions minimize global disturbance to the web. This framework extends Quine's insight by identifying truth as the stable fixed point of belief updating under reality's constraints. When a compression achieves strong lumpability (successful computational closure), further interaction with reality generates minimal error signals. The system has reached equilibrium: its internal model aligns with environmental structure closely enough that prediction errors approach zero. In Quine's terms, global disturbance is minimized: in our terms, systemic brittleness is minimized. The optimal constraint configuration thus represents the set of compressions that, once discovered, resist further revision not because we dogmatically cling to them but because reality's constraint structure leaves no gradient pointing elsewhere. Quine located truth in coherence plus minimal disturbance; this framework grounds that immanent conception in the thermodynamic necessity of computational closure.

7.3 Convergence as Information-Geometric Necessity

Why Different Cultures Converge on Similar Truths:

Not because of: - Shared biology alone (though this constrains) (Campbell 1974) - Social agreement (though this accelerates) (Longino 1990) - Divine revelation - Platonic access

But because: - Same constraint structure generates same compression optima - Independent ε-machine explorers facing identical landscape - Selection pressure eliminates high-brittleness compressions (Bradie 1986) - Thermodynamic attractors in the space of possible blanket configurations

Mathematical Analogy: Just as π appears to be discovered independently (Babylonians, Greeks, Indians, Chinese) because Euclidean constraints seem to determine it, scientific truths may be discovered independently because physical constraints appear to determine them.

Pluralism at the Frontier: Multiple viable compressions may exist (Pluralist Frontier), but catastrophically brittle ones (Negative Canon) are eliminated across all cultures.

Having developed the theoretical framework of computational closure and the Apex Network, we now apply these concepts to macro-epistemology, demonstrating how brittleness metrics provide practical tools for assessing the health of real-world knowledge systems and institutions.

Having established the optimal constraint configuration as the thermodynamic attractor for optimal compression, we now explore applications to macro-epistemology. The framework's information-theoretic foundation provides diagnostic tools for assessing systemic brittleness in knowledge systems, extending the analysis from individual cognition to collective epistemic health.

8. Applications: From Epistemology to Real-World Systems

8.1 Brittleness Metrics as Information Leakage Measures

These brittleness metrics formalize the accumulated cost of misalignment between a system's internal compressions and reality's constraint structure. Each metric tracks a different signature of prediction error: $P(t)$ measures how frequently predictions fail (requiring patches), $M(t)$ measures how bloated the model becomes in attempting to maintain failed compressions, $R(t)$ measures how few independent streams validate the compression, and $C(t)$ measures the energy spent suppressing error signals. Together, they quantify what Quine called "global disturbance"—the systemic strain when beliefs conflict with experience.

$P(t)$ (Patch Velocity): - Information-theoretic: Rate of local compression failures requiring ad-hoc fixes - Mechanistic: Blanket porosity increasing, closure failing - Phenomenology: Constant "but wait…" moments as predictions fail

$M(t)$ (Model Complexity): - Information-theoretic: Compression efficiency decreasing (more parameters, worse predictions) - Mechanistic: Failed lumpability forcing micro-tracking - **Clarification:** This tracks *Reactive Complexity* (ad-hoc patches, epicycles), not the *Irreducible Complexity* of accurate detailed models (like the Standard Model). Brittleness is indicated by the *growth* of complexity (dM/dt) in response to anomalies, rather than high absolute complexity.

$R(t)$ (Resilience Reserve): - Information-theoretic: Number of independent information streams successfully compressed - Mechanistic: Breadth of computational closure across domains - Phenomenology: Confidence from multi-source convergence

8.2 Coercion as Information Blindness

$C(t)$ (Coercive Overhead) deserves special attention:

Information-theoretic: Energy spent suppressing disconfirming information (creates information blindness)

Mechanistic: Maintaining rigid blanket against thermodynamic gradient while severing the error signal

Critical insight: Coercion is not just energetically costly but epistemically catastrophic. It eliminates the feedback loop needed to update the Markov blanket. By suppressing dissent (the primary data stream signaling misalignment), the system goes blind to reality's gradient, guaranteeing eventual collapse regardless of available resources.

The Thermodynamic Basis: This is not metaphor. Landauer's Principle establishes that information processing has irreducible physical cost floors. While the energy cost of erasing a single bit ($kT \ln 2$) is negligible at macroscopic scales, its implications for information processing become significant when considering the cumulative costs of maintaining complex, dynamic internal models over time. High information leakage (persistent prediction error) implies tangible metabolic and economic costs to any physical system. The link between epistemic brittleness and physical inefficiency is therefore functional, not merely metaphorical. When we speak of "information costs" in cognitive systems, we refer to resource constraints that have, at bottom, a thermodynamic grounding.

Viability versus Persistence: Critics may object that coercive systems often persist for decades or even millennia (e.g., dynastic Egypt, feudalism). This objection conflates **Static Durability** (the ability to resist change) with **Dynamic Viability** (the ability to adapt to change). Coercive systems maximize static durability by suppressing error signals ($C(t)$), effectively freezing the system state. This creates high stability as long as the environment remains constant. However, when the environment shifts, these systems lack the feedback loops to adapt, leading to catastrophic rather than gradual failure. As we detail in Section 9.4, this is "parasitic endurance"—metastability achieved by burning resources to mask brittleness, distinct from the structural viability of adaptive systems.

Phenomenology: Effortful belief maintenance ("I must avoid thinking about X"), defensiveness when challenged

8.3 The Real-World Necessity of Wholes

Reducing reality exclusively to particles isn't just metaphysically arid; it carries significant explanatory risks in practice. When we ignore computational closure and attempt to reason locally about parts, we fail to predict outcomes that are determined by boundary constraints.

Medicine (Atrial Fibrillation): Cardiologists do not treat atrial fibrillation by attempting to influence individual cardiomyocytes. The condition is an organ-level dynamical breakdown: a "rotor" or chaotic wave front. Effective treatment involves ablation, which alters the geometry of the heart's electrical boundary. The heart's macro-properties possess causal powers (sustaining or breaking the wave) that individual cells lack. The "whole" (the organ's geometry) is the causally effective intervention point.

Ecology (Salmon Collapse): Efforts to save salmon populations by focusing solely on the fish (hatcheries) often fail. The collapse is frequently a system-level failure involving river flow patterns, temperature gradients, and predator-balance. Restoring the population requires restoring the ecosystem's boundaries. The system possesses emergent stability properties invisible at the species level.

Technology (Internet Robustness): The internet exhibits traffic patterns and robustness properties that individual routers do not possess. Optimizing individual routers while ignoring network-level properties (like scale-free

topology) can lead to cascading failures. The network is a computational machine with its own closure; treating it as a pile of routers is an informational error.

Quantum Physics (Entanglement): The challenge "Does entanglement not violate locality?" is resolved by recognizing that locality holds at the macroscopic scales where wholes emerge. Even if the substrate allows non-local correlations, the statistical boundaries that define objects (like lab equipment or observers) preserve effective locality. Our argument applies where objects exist: above the quantum decoherence threshold where computational closure becomes possible.

8.4 Pragmatic Pushback as Thermodynamic Necessity

Information Can't Be Suppressed Indefinitely: Systems that maintain blankets misaligned with reality face thermodynamic costs: 1. Prediction errors accumulate (free energy increases) 2. Actions based on false compressions fail 3. Resources wasted compensating for misalignment 4. Eventually: Catastrophic collapse or forced revision

This Is Not Social Construction: It's physics. A bridge designed with false material-strength compressions will collapse regardless of social consensus.

The Ratchet Effect: Once a better compression is found (lower brittleness), reverting becomes thermodynamically unfavorable; you'd have to re-pay all the information costs the compression solved.

8.5 The Negative Canon as Compression Failure Archive

Every entry in the Negative Canon represents a failed computational closure: - Phlogiston: Combustion blanket leaked - Miasma: Disease blanket leaked - Lamarckian Inheritance: Evolution blanket leaked - Luminiferous Aether: Light-propagation blanket leaked

Educational Value: Studying failed compressions teaches the shape of constraint space: where the cliff edges are in the landscape of viable blankets.

Having applied the framework to macro-epistemology, we now consider broader implications for philosophy of mind, ethics, and metaphysics.

Having applied the framework to macro-epistemology and systemic brittleness, we now examine broader implications and remaining challenges. This framework offers a unified account of mind, knowledge, and truth, but its scope and limitations require careful consideration.

9. Implications and Open Questions

9.1 For Philosophy of Mind

A Naturalistic Framework (With Limits):

This framework offers a functional account of consciousness that identifies relevant distinctions without claiming to eliminate the hard problem:

- Phenomenology: May relate to what structural pattern recognition feels like from inside

- Conscious vs. Unconscious: Maps roughly onto structural vs. statistical pattern processing
- Qualia: Potentially compression gradients rendered in subjective space, though why these have qualitative character remains unexplained
- Self-awareness: Meta-blanket formation (blanket monitoring its own blankets)
- Unity of consciousness: Integrated information across blanket hierarchies
- *Institutional Minds Without Phenomenology*: While social institutions possess Markov blankets and exhibit information processing (they minimize surprise, maintain boundaries, adapt to environments), this framework does not imply they possess phenomenal consciousness. The difference lies in the timescale of integration. Human phenomenology relies on the immediate temporal binding of diverse information streams into a unified present moment (James 1890, 609). Institutional information processing occurs over days or months, far too slow for the integrated temporal binding characteristic of subjective experience. They are functional agents without phenomenology.

The Explanatory Gap Narrows But Doesn't Close: We've identified a functional distinction (statistical vs. structural pattern recognition) that appears to track the phenomenological boundary. This narrows the explanatory target: not "why does any information processing feel like something?" but "why does detecting mutual constraints feel like something?" That's progress, but the hard problem: why any functional process has subjective character, remains open.

9.2 For Epistemology

Knowledge Redefined: - Not justified true belief (Gettier problems) - But optimized compression validated by low brittleness

Justification Naturalized: - Internal coherence (Level 3) necessary but insufficient - External validation (pragmatic testing) required - Truth tracks optimal compression, not correspondence to pre-existing propositions

This framework naturalizes Quine's epistemological holism. Beliefs do not stand alone as isolated justified-true-belief atoms but form an interconnected web where revision in one area propagates throughout the system. The brittleness metrics formalize what Quine described informally as "global disturbance"—the cost of revising beliefs when they conflict with experience. Central beliefs resist revision not because they are metaphysically privileged but because changing them requires reorganizing vast portions of the web. Truth, in this framework, extends Quine's pragmatist insight: it is not static correspondence to Platonic propositions but the stable configuration that minimizes brittleness under reality's constraints: what Quine would recognize as the limit of inquiry where further revisions generate more disturbance than they resolve.

9.3 For Metaphysics

Ontology Enacted, Not Discovered: - What "exists" = what blankets successfully compress - Different blanket configurations = different ontologies - But not arbitrary: reality constrains which blankets close

Emergence Mechanized: - New causal levels arise from successful computational closure - Not mysterious or epiphenomenal but thermodynamically real - The universe is layered compression hierarchies all the way up

9.4 For Ethics: The Thermodynamic Costs of Coercion

The Markov Blanket View of Moral Agency: If generative priors are successful Markov blankets, then ethics concerns how we draw boundaries around the agency of others.

Methodological Note (The Is/Ought Boundary):

We must acknowledge a crucial limit. This framework describes how certain social configurations generate higher or lower thermodynamic costs. However, it does not (and cannot) directly derive moral obligations from these descriptive facts. The move from "X generates high brittleness" to "therefore, X is morally wrong" crosses Hume's is/ought gap. (For a detailed treatment of this procedural model, see Glenn 2025).

What we can legitimately claim: Certain moral intuitions may have information-theoretic grounding. The recognition that denying others' agency generates catastrophic social costs helps explain why such behaviors are unsustainable. It also helps explain why moral progress often tracks toward lower-brittleness configurations.

But whether we should care about minimizing brittleness, or whether thermodynamic efficiency has moral relevance, remains a normative question this framework doesn't fully resolve.

Two-Part Defense of the Framework's Scope:

The Constitutive Defense: The framework's apparent focus on "persistence" as a value warrants clarification. Persistence functions as a methodological constraint, not a normative commitment. Any normative system that fails to persist becomes unavailable for historical analysis and comparative study. We describe the filtering process that operates on persistent systems without claiming persistence is intrinsically valuable or morally good. The framework maps which configurations enable persistence under pragmatic constraints, not which configurations ought to exist. This is analogous to how evolutionary biology describes which traits enable reproductive success without claiming reproduction is morally good—persistence is the entry condition for having a track record to analyze, not a value we endorse.

The Instrumental Defense: The framework offers conditional oughts, not categorical imperatives. The implicit hypothetical imperative is: "If a system aims to persist while enabling cooperation and minimizing coordination costs, then it should adopt low-brittleness principles and avoid configurations that generate catastrophic thermodynamic costs." For systems indifferent to persistence, or for evaluating dimensions of value beyond viability (aesthetic worth, spiritual meaning, individual dignity independent of systemic function), our framework provides no normative force. We claim only that thermodynamic viability is necessary for persistence, not that it is sufficient for moral worth or that it exhausts moral considerations.

Independence of Pragmatic Constraints: A deeper worry concerns circularity—are we simply reading normative conclusions off the constraints we happen to measure? This misunderstands the relationship between the filtering process and the constraints that do the filtering. The pragmatic constraints that filter normative systems—biological needs (caloric requirements, immune function, reproductive capacity), cognitive limitations (bounded rationality, working memory constraints, information processing capacity), coordination requirements (game-theoretic necessities of cooperation under potential defection,

common-pool resource management), and physical necessities (thermodynamic imperatives, resource scarcity)—are empirically discoverable facts about human embodiment and social organization, not value judgments or normative commitments. These constraints are determined by biology, physics, psychology, and mathematics. They are what filter normative systems through differential brittleness; they are not products of the filtering process itself. A society committed to asceticism still faces biological caloric requirements. A society valuing absolute equality still faces coordination constraints on large-scale cooperation. These are mind-independent constraints discovered through standard empirical inquiry.

Floor vs. Ceiling Clarification: This framework maps normativity's floor (non-negotiable viability conditions required for any persistent cooperative system), not its ceiling (the full space of human flourishing, aesthetic value, spiritual meaning, or dignity). Societies must secure the floor—avoiding catastrophic failures like endemic starvation, demographic collapse, or coordination breakdown—before pursuing higher goals. The framework provides no guidance on which of many viable configurations is best, most beautiful, most meaningful, or most just by dimensions orthogonal to thermodynamic viability. It identifies what cannot work sustainably, not what ought to be pursued among viable options. Moral philosophy beyond viability constraints requires additional normative frameworks this paper does not provide.

With these qualifications in place, we can explore how the information-theoretic perspective illuminates ethical phenomena without claiming to have derived ethics from thermodynamics. This methodological boundary doesn't invalidate the descriptive analysis that follows; it clarifies the limits of what the framework can legitimately claim.

Information-Theoretic Analysis of Agency Denial:

Rosas et al. (2024) demonstrate that causally closed systems can be efficiently controlled through macro-level interventions—engaging with their computational closure rather than manipulating their substrate. This insight provides a mechanistic account of moral interaction: when we engage with another agent's reasons, beliefs, and goals (their ε-machine), we interact efficiently with the causally autonomous level. We work through their computational closure, allowing their internal dynamics to determine outcomes. When we bypass their agency to force their physical body or manipulate their circumstances (intervening on the substrate), we breach their causal closure and must manage all the micro-level resistance their autonomous system generates.

This explains why persuasion is thermodynamically cheaper than coercion: persuasion operates at the ε-machine level (engaging autonomous macro-dynamics), while coercion operates at the substrate level (fighting against those dynamics). Substrate interference is thermodynamically expensive in a specific way: by discarding the efficient causal structure of the agent's computational closure, the coercer must manage the full complexity of the micro-states directly. Persuasion leverages the agent's own compression architecture; coercion fights against it. The computational closure Rosas et al. identify is precisely what moral recognition respects and evil violations ignore.

Evil as Closure Breach (Bypassing the ε-Machine to Manipulate the Substrate):

When a system (individual, institution, ideology) treats other agents as mere objects (as parts of the external environment to be manipulated rather than as

causally closed entities with autonomous ε-machines), it commits a specific information-theoretic error:

Failed Closure Recognition: - Moral agents: Achieve computational closure (autonomous ε-machines, internal goals, reactive capacities) - Objects: Lack computational closure (can be freely manipulated without resistance) - Evil: Treating agents as objects (closure breach, forcing the substrate rather than engaging the software)

Thermodynamic Consequences:

Moral Configuration	Information Structure	Brittleness Signature
Recognition of Closure	Engaging with others' ε-machines → arguments, persuasion, negotiation work through their computational closure	Low $C(t)$: Coordination via understanding; Low $P(t)$: Predictable responses when modeling their goals/beliefs
Breach of Closure	Bypassing ε-machines to manipulate substrate → coercion, violence, deception force the hardware while ignoring the software	High $C(t)$: Massive coercion needed to suppress autonomous ε-machine responses; High $P(t)$: Constant resistance as closed systems fight substrate manipulation

Why Closure Breach Generates Brittleness: When you bypass an agent's ε-machine (their will, reasoning, goals) to force their substrate (their body, circumstances), you lose the predictive benefits of their internal model. You must now manage every micro-variable yourself, constantly suppressing their autonomous responses. The agent's computational closure actively resists your interventions, generating persistent prediction errors and requiring escalating coercion costs.

The key mechanism is parasitic endurance: these systems survive not through structural viability but by extracting surplus energy to pay the massive coercion costs. They appear stable only because they burn external resources—extracting surplus from the oppressed, exploiting resource windfalls (oil wealth, foreign aid), or cannibalizing their own future—to mask the entropy generated by their internal friction. This is energetic insolvency maintained through extraction, not genuine thermodynamic efficiency.

Metastability Through Information Cannibalization: How do high-brittleness systems like totalitarian regimes persist? They engage in *information cannibalization*. Viable systems invest energy in sensors (dissent, market prices, free speech) to update their Markov blankets. Coercive systems consume that sensory infrastructure to fuel immediate stability. By destroying the error-generating mechanisms (purging experts, fixing prices, censoring press), they artificially lower free energy in the short term (internal surprise is minimized because no one dares report bad news).

This creates what we call Zombie Stability. Just as a starving body consumes its own muscle tissue for quick energy, a coercive state consumes its own truth-finding mechanisms (dissent, journalism, expertise) to fuel immediate stability. It trades long-term adaptability (muscle) for short-term survival (sugar). The system looks robust but has eaten the very organs required to detect and adapt to future change. This clarifies the normative judgment: such systems are "evil"

not merely because they cause pain, but because they are engaged in the systemic consumption of their own resilience reserve, $R(t)$. They trade the capacity to adapt to the future for the capacity to control the present, a thermodynamic strategy that is mathematically guaranteed to terminate in catastrophic failure.

The claim is not that coercive systems fail instantly, nor that thermodynamic efficiency is identical to moral goodness. It is possible, in principle, to conceive of a system that minimizes internal free energy by engineering out the capacity for agency entirely—a "Brave New World" scenario where the subject population loses the capacity to generate prediction errors. While such a system might achieve low internal friction (low $C(t)$), it would maximize external brittleness by eliminating the diversity of epistemic sensors required to detect novel environmental constraints.

By suppressing the variation engine of individual agency, such systems reduce their Resilience Reserve ($R(t)$) to near zero. They are optimized for a static present but structurally guaranteed to collapse in a dynamic future. Thus, the thermodynamic critique of coercion is not that it is morally wrong per se, but that it creates a specific form of fragility: it purchases short-term stability at the price of long-term adaptability. The alignment between moral intuition and thermodynamic viability emerges because the conditions required for long-term epistemic resilience—autonomy, diversity, and open feedback loops—overlap significantly with the conditions of moral recognition.

These systems are metastable, not viable: appearing stable only while the subsidy lasts. Eventually, when external resources are exhausted or resistance accumulates beyond suppression capacity, the thermodynamic gradient asserts itself. However, this process can span centuries. This is a long-run structural constraint, not a guarantee of swift moral justice.

Systems that refuse to engage with others' computational closure accumulate elevated brittleness through parasitic endurance: - Slavery: Survived not through structural stability but by extracting surplus labor to pay the massive enforcement costs ($C(t)$), though this created persistent resistance ($P(t)$) and eventually collapsed when extraction could no longer cover coercion costs - Totalitarianism: Achieves "zombie stability" by burning external resources (oil revenue, international subsidies, or the cannibalization of internal capital) to fund surveillance states while denying citizens' agency; thermodynamically insolvent but can maintain metastability while resources last - Genocide: Ultimate closure denial (erasing agents entirely when the energetic cost of modeling their agency becomes perceived as exceeding the extraction capacity)

Not Moral Relativism: Different cultures may draw different boundaries around "who counts as an agent" (children? animals? ecosystems?), but systems that catastrophically misalign with the actual distribution of agency in their environment pay thermodynamic costs. The optimal constraint configuration includes the recognition that other humans are agents: not because of moral axioms but because any other blanket configuration generates unsustainable brittleness.

Connection to Expansion of Moral Circle: Historical moral progress often involves recognizing previously-objectified groups as agents (women, slaves, colonized peoples). This isn't just "being nicer"; it's discovering that modeling these groups as having Markov blankets dramatically reduces social brittleness (abolishing slavery eliminates massive $C(t)$ costs of enforcement).

9.5 Open Challenges

The Integration Problem: How do separate blankets integrate into unified experience? What determines which compressions "bind" into single qualia?

The Novelty Problem: How do genuinely new compressions arise? Is all creativity just recombination, or can systems generate truly novel blanket configurations?

The Value Problem (Partially Addressed): Section 9.4 shows how evil can be understood as high-entropy sociology: denying others' Markov blankets. But open questions remain: Can all moral truths be reduced to thermodynamic efficiency? What about irreducibly normative dimensions (beauty, meaning, sacred values) that resist compression-theoretic analysis? These questions require further development beyond This paper's scope.

The Limits Problem: Are there hard limits to what can be compressed? Gödel's incompleteness theorems suggest some truths resist finite compression. Quantum mechanics presents another potential limit case, where classical structural constraints may not apply straightforwardly—the framework assumes causal relationships can be represented as mutual constraints between variables, but quantum entanglement and superposition challenge this classical architecture. Whether computational closure can be meaningfully extended to quantum systems, or whether quantum phenomena represent a fundamental boundary for blanket-based ontology, remains an open theoretical question. Exploring the implications for the optimal constraint configuration framework remains important future work.

10. Conclusion: A Naturalistic Framework (With Acknowledged Limits)

This paper has developed an information-theoretic framework connecting raw information processing to conscious awareness to objective truth through compression, blanket formation, and thermodynamic selection. These connections remain theoretical and require empirical validation.

The framework's core claims: Information is processed by all physical systems. Compression creates dispositions as minimal encodings of regularities. Markov blankets emerge when compressions achieve statistical boundaries. Computational closure succeeds when blankets create autonomous causal levels. Consciousness may relate to meta-blanket hierarchies and structural pattern recognition. Generative priors are culturally transmitted successful closures. Brittleness measures information leakage when closures fail. Pragmatic selection eliminates high-brittleness compressions.

Truth, on this account, is not correspondence to static propositions but alignment with the optimal constraint configuration: the state where a system's enacted boundaries map the environment's causal constraints, achieving maximal computational closure with minimal information leakage.

This framework employs information-theoretic language as conceptual scaffolding for understanding epistemic and cognitive phenomena. Whether brains literally implement variational free energy minimization or generative priors are literally encoded as Markov blankets in neural tissue remains an empirical question. The philosophical insights—about what makes predicates successful, how knowledge

systems fail, and why inquiry converges—retain their force even if specific mechanistic claims require revision, though the overall framework itself remains subject to empirical testing and philosophical refinement.

We've identified relevant functional distinctions without solving all foundational problems. The hard problem of consciousness (why structural pattern recognition feels like anything) remains open. The is/ought gap, or why thermodynamic efficiency should matter morally, is not derived from the framework. The plurality question, concerning how much genuine pluralism exists versus forced convergence, remains empirical. Finally, the limits of compression, such as whether some truths resist finite compression due to incompleteness theorems, are not fully addressed.

This framework provides a mechanistic account of how notions form and validate through statistical versus structural pattern recognition. It explains why singular experiences can carry immediate epistemic weight and offers a naturalistic grounding for cross-cultural convergence in knowledge systems. It also delivers a functional account of consciousness that narrows the explanatory gap, a framework for understanding truth as constraint-determined structure rather than correspondence to Platonic forms, and an information-theoretic analysis of why coercion generates systemic brittleness.

Consciousness, agency, and truth remain real within this framework: not eliminated or reduced away, but understood as emerging from information processing under constraint. The framework is naturalistic without being eliminativist.

Glenn (2025) develops these concepts in application to macro-epistemology and historical knowledge systems. This paper provides the theoretical grounding in information theory, computational closure, and constraint-driven selection.

Ultimately, this framework suggests a unification of metaphysics and information theory. Reality is not a flat soup of particles where only the smallest things are real. It is a lattice of computational machines, nested within one another. Each machine (a cell, a mind, a society) maintains its own closure and runs its own causal code. Science is not the reduction of these machines to their parts, but the discovery of which machine correctly predicts the phenomenon at hand.

This framework positions us not as passive observers of a pre-existing Platonic reality, but as active participants in discovering the constraint-determined structure of our universe. Knowledge advances through the systematic elimination of configurations that generate unsustainable brittleness, gradually mapping the optimal constraint configuration.

References

Aslin, Richard N., and Elissa L. Newport. 2012. "Statistical learning: From acquiring specific items to forming general rules." *Current Directions in Psychological Science* 21(3): 170–176. https://doi.org/10.1177/0963721412436806.

Ayvazov, Mahammad. 2025. "Toward a Phase Epistemology: Coherence, Response and the Vector of Mutual Uncertainty." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5250197.

Bennett, Charles H. 1988. “Logical Depth and Physical Complexity.” In *The Universal Turing Machine: A Half-Century Survey*, edited by Rolf Herken, 227–257. Oxford: Oxford University Press.

BonJour, Laurence. 1985. *The Structure of Empirical Knowledge*. Cambridge, MA: Harvard University Press. ISBN 978-0674843813.

Campbell, Donald T. 1974. “Evolutionary Epistemology.” In *The Philosophy of Karl R. Popper*, edited by Paul A. Schilpp, 413–63. La Salle, IL: Open Court.

Chalmers, David J. 2006. “Strong and Weak Emergence.” In *The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion*, edited by Philip Clayton and Paul Davies, 244–54. Oxford: Oxford University Press. ISBN 9780199287147. https://doi.org/10.1093/acprof:oso/9780199287147.003.0011.

Clark, Andy. 2013. “Whatever next? Predictive brains, situated agents, and the future of cognitive science.” *Behavioral and Brain Sciences* 36(3): 181–204. https://doi.org/10.1017/S0140525X12000477.

Dennett, Daniel C. 1991. “Real Patterns.” *Journal of Philosophy* 88(1): 27-51. https://doi.org/10.2307/2027085.

Dennett, Daniel C. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press. ISBN 978-0262540537.

Dittrich, Christian, and Jennifer Flygare Kinne. “The Information-Theoretic Imperative: Compression and the Epistemic Foundations of Intelligence.” Preprint, submitted October 30, 2024. arXiv:2510.25883 [cs.AI]. https://doi.org/10.48550/arXiv.2510.25883.

Friston, Karl J. 2010. “The Free-Energy Principle: A Unified Brain Theory?” *Nature Reviews Neuroscience* 11 (2): 127–138. https://doi.org/10.1038/nrn2787.

Friston, Karl J., Thomas Parr, Conor Heins, Axel Constant, Daniel Friedman, Takuya Isomura, Chris Fields, Tim Verbelen, Maxwell Ramstead, John Clippinger, and Christopher Frith. 2025. “What the Flock Knows That the Birds Do Not: Exploring the Emergence of Joint Agency in Multi-Agent Active Inference.” *arXiv* preprint arXiv:2511.10835. https://arxiv.org/abs/2511.10835.

Glenn, Patrick. 2025. “The Architecture of Failure: How Systemic Brittleness Drives Convergent Coherence to Forge Objective Truth.” PhilPapers. https://philpapers.org/rec/GLETAO.

James, William. 1890. *The Principles of Psychology*. 2 vols. New York: Henry Holt and Company.

Kirchhoff, Michael, Thomas Parr, Ensor Palacios, Karl Friston, and Julian Kiverstein. 2018. “The Markov Blankets of Life: Autonomy, Active Inference and the Free Energy Principle.” *Journal of the Royal Society Interface* 15(138): 20170792. https://doi.org/10.1098/rsif.2017.0792.

Ladyman, James, and Don Ross. 2007. *Every Thing Must Go: Metaphysics Naturalized*. Oxford: Oxford University Press.

Landauer, Rolf. 1961. “Irreversibility and Heat Generation in the Computing Process.” *IBM Journal of Research and Development* 5(3): 183–191. https://doi.org/10.1147/rd.1961.5.3.183. ISBN 978-0199276196.

Laukkonen, Ruben, et al. 2025. “A Beautiful Loop: An Active Inference Theory of Consciousness.” *Neuroscience & Biobehavioral Reviews* 176: 106296. https://doi.org/10.1016/j.neubiorev.2025.106296.

Longino, Helen E. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton, NJ: Princeton University Press. ISBN 978-0691020518.

Mangalam, Madhur. 2025. "The Emperor's New Pseudo-Theory: How the Free Energy Principle Ransacked Neuroscience." Preprint. DOI: 10.31234/osf.io/azkgc. https://osf.io/azkgc.

Marcus, Gary F. 2001. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge, MA: MIT Press. ISBN 978-0262632683.

Oizumi, Masafumi, et al. 2014. "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0." *PLOS Computational Biology* 10(5): e1003588. https://doi.org/10.1371/journal.pcbi.1003588.

Olsson, Erik J. 2005. *Against Coherence: Truth, Probability, and Justification*. Oxford: Oxford University Press. ISBN 978-0199279999.

Parr, Thomas, and Karl J. Friston. 2025. "How Do Inner Screens Enable Imaginative Experience? Applying the Free-Energy Principle to Attention." *Neuroscience of Consciousness* 2025(1): niaf009. https://doi.org/10.1093/nc/niaf009.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press. ISBN 978-0521773621.

Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: MIT Press. ISBN 978-0262670012.

Rosas, Fernando E., et al. 2024. "Disentangling High-Order Mechanisms and High-Order Behaviours in Complex Systems." *Nature Physics* 20: 1095–1104. https://doi.org/10.1038/s41567-024-02477-4.

Shannon, Claude E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27(3): 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Sinclair, Robert. 2007. "Quine's Naturalized Epistemology and the Third Dogma of Empiricism." *Southern Journal of Philosophy* 45, no. 3: 455–472. https://doi.org/10.1111/j.2041-6962.2007.tb00060.x.

Solms, Mark. 2019. "The Hard Problem of Consciousness and the Free Energy Principle." *Frontiers in Psychology* 9: 2714. https://doi.org/10.3389/fpsyg.2018.02714.