The Information-Theoretic Imperative:
Compression and the Epistemic Foundations of Intelligence

Christian Dittrich[1] and Jennifer Flygare Kinne[2]

[1]Sciquoia AG, Basel, Switzerland
[2]Faculty of Arts and Sciences, Harvard University, Cambridge, MA, USA

Abstract

Existing frameworks converge on the centrality of compression to intelligence but leave underspecified why this process enforces the discovery of causal structure rather than superficial statistical patterns. We introduce a two-level framework to address this gap. The Information-Theoretic Imperative (ITI) establishes that any system persisting in uncertain environments must minimize epistemic entropy through predictive compression: this is the evolutionary "why" linking survival pressure to information-processing demands. The Compression Efficiency Principle (CEP) specifies how efficient compression mechanically selects for generative, causal models through exception-accumulation dynamics, making reality alignment a consequence rather than a contingent achievement.

Together, ITI and CEP define a causal chain: from survival pressure to prediction necessity, compression requirement, efficiency optimization, generative structure discovery, and ultimately reality alignment. Each link follows from physical, information-theoretic, or evolutionary constraints, implying that intelligence is the mechanically necessary outcome of persistence in structured environments.

This framework yields empirically testable predictions: compression efficiency, measured as approach to the rate-distortion frontier, correlates with out-of-distribution generalization; exception-accumulation rates differentiate causal from correlational models; hierarchical systems exhibit increasing efficiency across abstraction layers; and biological systems demonstrate metabolic costs that track representational complexity. ITI/CEP thereby provides a unified account of convergence across biological, artificial, and multi-scale systems, addressing the epistemic and functional dimensions of intelligence without invoking assumptions about consciousness or subjective experience.

Keywords: Information theory; compression efficiency; predictive coding; epistemic alignment; intelligence; causality; minimum description length; free energy principle

1 Introduction

Over the past several decades, a striking convergence has emerged across artificial intelligence, neuroscience, and theoretical biology: systems that learn, predict, or survive most effectively are those that compress most efficiently. The principles of the Information Bottleneck are now understood to govern aspects of deep learning optimization (Tishby and Zaslavsky, 2015; Shwartz-Ziv and Tishby, 2017; Saxe et al., 2019). Recent work shows that the remarkable generalization capabilities of Large Language Models (LLMs) are deeply tied to their success in compressing vast textual data, mechanically rediscovering the generative structure embedded within (Li et al., 2024). In neuroscience, predictive coding is framed as a form of hierarchical compression ((Rao and Ballard, 1999; Huang and Rao, 2011), providing a plausible implementation of Karl Friston's Free Energy Principle, which casts biological intelligence as the minimization of surprise through efficient prediction (Friston, 2010; Buckley et al., 2017). The Minimum Description Length (MDL) principle formalizes Occam's razor as an optimization of compression (Rissanen, 1978; Grünwald, 2007). Jürgen Schmidhuber's theory of compression progress explains curiosity and creativity as rewards for improved compressibility (Schmidhuber, 1991, 2010). Recent work continues to explore these connections across domains (Bengio and Malkin, 2024; Clark, 2020; Lake et al., 2017) with renewed interest in compression-based intrinsic motivation (Achiam and Sastry, 2017).

These frameworks, while converging on the centrality of compression, leave a critical mechanism underspecified. How does compression efficiency translate into epistemic validity? That is, how does minimizing a model's description length come to entail discovering the real causal structure of the world, rather than an arbitrary but compact fiction? This question addresses the fundamental puzzle of epistemic alignment, why systems optimizing purely information-theoretic objectives converge on models that track actual environmental regularities. This is a topic of central importance in both philosophy of science (Woodward, 2003; Cartwright, 2007) and AI safety (Amodei et al., 2016; Patankar et al., 2023).

We propose a two-level framework to fill this gap, providing the missing mechanistic link between compression and causality. The Information-Theoretic Imperative (ITI) establishes why systems must compress. It posits that any system persisting in an uncertain environment must minimize epistemic entropy – the remaining uncertainty within a system's internal model about what future states will occur – through predictive compression. This is the evolutionary and thermodynamic driver connecting survival pressure to universal information-processing requirements. The Compression Efficiency Principle (CEP) specifies how this process enforces reality-alignment. It states that efficient compression mechanically selects for generative, causal models over superficial correlations. This is the architectural mechanism that makes alignment with reality a consequence of optimization, not a contingent achievement. The synthesis of these principles unfolds as a causal chain, where each link follows from physical, information-theoretic, or evolutionary constraints:

$$\text{Survival Pressure} \rightarrow \text{Prediction Necessity} \rightarrow \text{Compression Requirement}$$
$$\rightarrow \text{Efficiency Optimization} \rightarrow \text{Generative Structure Discovery} \rightarrow \text{Reality Alignment}$$

This chain makes intelligence not a special biological accomplishment but the mechanically necessary outcome of persistence in a structured environment. It explains phenomena ranging from the architecture of evolved neural systems to the emergence of reality-aligned reasoning in artificial systems trained purely on statistical patterns.

Contributions and Differentiation   Our framework provides three core contributions that distinguish it from existing theories:

First, we establish the causal mechanism linking compression efficiency to epistemic validity. While MDL shows that minimal descriptions favor simpler models, it doesn't explain why simplicity tracks truth. We demonstrate that superficial patterns accumulate exceptions that inflate description length, while generative

models encode the mechanisms that produce variation, achieving constant structural cost amortized across instances.

Second, we connect survival imperatives to architectural constraints at the universal level. The FEP shows that biological systems minimize surprise through prediction (Friston, 2010), and Schmidhuber's theory explains learning rewards via compression progress (Schmidhuber, 2009, 2010). ITI/CEP extends these insights by demonstrating that persistence requirements propagate through information-theoretic necessity to compression-efficiency constraints that enforce the discovery of reality. This mechanism operates beyond single-agent learning dynamics or specific biological substrates, explaining why diverse systems – from neural networks to evolutionary processes – converge on reality-testing architectures.

Third, we generate a set of testable predictions that distinguish our framework empirically. We predict that compression efficiency (measured as an approach to the rate-distortion frontier) correlates with out-of-distribution generalization, that exception accumulation rates differentiate causal from correlational models, and that hierarchical compression amplifies efficiency across layers of abstraction. These predictions allow for empirical validation that moves beyond conceptual synthesis.

We position ITI/CEP as a complementary framework, not a replacement. It shares the FEP's survival-to-prediction logic while adding the specific mechanism of compression efficiency. It builds on MDL's parsimony principle (Rissanen, 1978) while explaining why minimum description discovers generative structure. It extends Schmidhuber's concept of compression progress from an agent-level learning reward to a universal epistemic constraint.

Scope and Limitations   ITI/CEP is an epistemic-functional framework: it explains how systems extract and validate environmental structure through compression and prediction. We make no claims about consciousness, subjective experience, or phenomenology. The framework explains what constraints intelligent systems face and what mechanisms enforce reality-alignment, not the complete architecture of any particular implementation. Mathematical formalism is developed where established foundations exist, such as Information Bottleneck efficiency and rate-distortion theory, while flagging areas that still require formal development, such as the precise rates of exception accumulation.

The framework remains falsifiable: it predicts its own refinement through the discovery of exceptions. Persistent deviations from the efficiency–causality link would signal compression failure – evidence that the system's exception dynamics have not yet converged. In this way, ITI/CEP treats falsification itself as a continuation of learning: each discovered inconsistency tightens the bound between efficiency and truth.

A Note on Methodological Self-Consistency   We prioritize conceptual clarity and empirical tractability over premature formalization. This approach follows from the framework's own epistemology: this framework treats all knowledge (including formal theory) as compression of environmental regularities at different signal-to-noise ratios. We provide formal lemmas (Appendix A) not because formalism holds privileged epistemic status, but because symbolic precision improves predictive testability and communicability. Formalism serves prediction; prediction tests compression; compression reveals structure. The same physical logic applies to the framework itself.

This perspective has methodological consequences. The framework validates not through mathematical closure but through convergent pattern recognition across independent domains – the criterion it predicts all theories face. Sections 6-7 show how this applies.

Chapter Organization   Section 2 reviews related work across information theory, predictive frameworks, algorithmic curiosity, and complexity science, identifying the gap our framework addresses. Section 3 develops the Information-Theoretic Imperative: the universal requirement that persistent systems minimize epistemic entropy. Section 4 presents the Compression Efficiency Principle: the mechanism by which

efficient compression enforces generative structure discovery. Section 5 synthesizes these levels into a unified causal chain from survival to reality-alignment. Section 6 provides testable predictions enabling empirical validation. Section 7 discusses scope, limitations, and relationships to existing frameworks. Section 8 concludes with implications for understanding intelligence across biological, artificial, and multi-scale systems, illustrating how the same information-theoretic constraints may operate from molecular to collective levels of organization.

Ultimately, the ITI/CEP framework reveals that compression is not mere computational convenience but a physical necessity. Information-theoretic constraints on efficiency become the very mechanism that enforces the epistemic alignment we recognize as intelligence.

2 Related Work

This section situates the Information-Theoretic Imperative (ITI) and Compression Efficiency Principle (CEP) within prior work spanning information theory, neuroscience, and complexity science, outlining the conceptual lineage from which our framework extends.

2.1 Information-Theoretic Foundations

The idea that intelligence is a form of information processing has deep roots. Shannon's foundational work established entropy as the measure of uncertainty, defining information as that which reduces it (Shannon, 1948). Independently, Kolmogorov and Solomonoff developed algorithmic information theory, defining the complexity of an object as the length of the shortest program that generates it (Kolmogorov, 1965; Solomonoff, 1964). This formalization suggested a profound connection: to understand is to compress. A phenomenon is understood to the degree that it can be described concisely; irregular patterns resist compression, while lawful regularities permit it.

The Minimum Description Length (MDL) principle operationalizes this insight for statistical inference. Formalizing Occam's razor, MDL posits that the best model is the one that minimizes the total description length: the sum of the model's own complexity and the length of the data encoded with that model's help (Rissanen, 1978; Grünwald, 2007). This provides a rigorous criterion for model selection that has proven to be remarkably successful in domains ranging from neural coding (Balasubramanian, 1997) to causal discovery (Janzing and Schölkopf, 2010).

The Information Bottleneck (IB) framework extends these ideas to predictive systems. Tishby and colleagues formalized the trade-off between compression and prediction: an optimal representation compresses an input variable X maximally while preserving the maximum possible information about an output variable Y (Tishby et al., 1999). This rate-distortion perspective reveals compression not merely as a tool for efficiency but as a fundamental principle for extracting meaningful structure. Recent work demonstrates that deep learning implicitly implements IB principles, with neural network layers progressively compressing training data while retaining predictive power (Tishby and Zaslavsky, 2015; Shwartz-Ziv and Tishby, 2017; Saxe et al., 2019). Extensions to IB formalize compression's role in causal discovery: compression under deterministic dynamics selects causal over spurious patterns (Strouse and Schwab, 2016), while optimal causal inference emerges from minimizing stored information versus predictive power (Still et al., 2010).

These frameworks establish the computational link between compression and understanding. They specify what optimal systems should do but do not fully explain why compression-optimized systems should converge on reality-aligned representations instead of any other compressible fiction.

2.2 Predictive and Energetic Frameworks

A parallel line of inquiry in neuroscience and thermodynamics establishes the survival-based imperative for prediction and efficiency.

Friston's Free Energy Principle (FEP) provides a complementary perspective grounded in the physics of self-organization (Friston, 2010; Friston and Stephan, 2007). The FEP posits that any system that maintains its states in a bounded region far from thermodynamic equilibrium must minimize its variational free energy; an upper bound on surprise (the negative log-likelihood of sensory observations). This imperative unifies perception and action: organisms minimize prediction error both by updating their internal models of the world (perception) and by acting to change the world to make it more predictable (action). The FEP has generated substantial empirical support, explaining phenomena from sensory attenuation to active inference (Friston et al., 2017; Parr et al., 2020; Da Costa et al., 2021).

Predictive coding offers a plausible neural architecture for implementing the FEP's computational goals (Rao and Ballard, 1999; Friston and Kiebel, 2009; Huang and Rao, 2011). In this model, higher levels of a cortical hierarchy predict the activity of lower levels, with only the residual prediction errors being propagated upward for model updating. This accounts for a diverse range of perceptual and cognitive phenomena through a single, elegant principle: brains are prediction machines that minimize surprise through hierarchical inference (Clark, 2016; Hohwy, 2013).

The efficient coding hypothesis extends this predictive view to the level of sensory neurons (Barlow, 1961; Simoncelli and Olshausen, 2001). It suggests that sensory systems should encode environmental statistics as efficiently as possible, maximizing information transmission under strict metabolic constraints. This hypothesis correctly predicted specific neural coding properties that were subsequently confirmed experimentally (Olshausen and Field, 1996; Laughlin, 2001).

Together, these frameworks establish that prediction and efficiency are central to both neural function and organismal survival. They do not, however, fully specify the mechanism by which efficient prediction enforces the manifestation of causal rather than merely correlational structure.

2.3 Algorithmic Curiosity and Compression Progress

While the FEP provides a top-down, systemic account, a third line of research examines the bottom-up, algorithmic drivers of learning and curiosity, again finding compression at its core.

Schmidhuber's work on compression progress provides the most direct precursor to our framework (Schmidhuber, 1997a,b, 2006, 2009, 2010). This theory posits that learning systems are intrinsically motivated by improvements in their ability to compress their sensory history. An agent's "curiosity reward" is proportional to the first derivative of its compression performance. In other words, learning progress itself is reinforcing. This framework elegantly explains a wide range of phenomena, from artistic beauty and humor to scientific discovery, as manifestations of a fundamental drive toward improved compressibility. By implementing these principles in artificial agents, Schmidhuber demonstrated genuine curiosity-driven exploration, providing an algorithmic account of developmental learning and creativity grounded in information theory.

The compression progress framework operates primarily at the agent level, specifying the reward signals that drive learning. It explains what motivates an intelligent system to improve its models. It does not, however, fully explain why efficient compression itself – independent of progress – should enforce the discovery of reality-aligned structure.

2.4 Complexity and Adaptive Systems

A fourth perspective comes from the Santa Fe Institute tradition, which frames intelligence as an emergent property of complex adaptive systems (Mitchell, 1998, 2009; Flack and Krakauer, 2011; Krakauer et al., 2020). This view emphasizes information processing as fundamental to life and mind. Krakauer and colleagues define intelligence as systems that process environmental information to guide adaptive behavior, encompassing biological organisms and artificial systems (Krakauer et al., 2020).

Mitchell's work shows how distributed computation can yield coordinated behavior without centralized control. Building on this foundation, Krakauer's information theory of individuality defines organisms as entities that preserve informational dependencies across time, grounding individuality in the ability to compress past states into present structure – establishing that even the boundaries between self and environment emerge from compression dynamics. Flack and Krakauer then extend this reasoning to collective systems, demonstrating how groups compress environmental uncertainty into shared predictive structure, revealing compression as a principle operating across scales from individual organisms to social collectives.

This complexity perspective illuminates how adaptive information processing can yield coherent behavior under uncertainty, yet it stops short of specifying why compression itself enforces epistemic validity – the

mechanism addressed by the ITI/CEP framework.

2.5 The Gap

These diverse frameworks establish several key principles:

- Understanding requires compression (MDL, IB)

- Survival requires prediction (FEP, efficient coding)

- Compression progress drives learning (Schmidhuber)

- Intelligence emerges from information processing under constraints (complexity science)

Yet a critical mechanism remains underspecified: Why does efficient compression mechanically enforce the appearance of generative, causal structure rather than superficial statistical patterns?

MDL shows that minimal descriptions favor simpler models but does not explain why simplicity tracks truth (Cartwright, 2007). The FEP shows that survival requires prediction but does not detail how this requirement enforces reality-alignment over useful fictions. Schmidhuber's theory explains learning motivation but treats compression quality as a given rather than explaining what makes a compression scheme efficient in the first place. The complexity tradition identifies information processing as fundamental, but does not formalize the link between compression constraints and epistemic accuracy.

The missing piece is a causal account of how compression efficiency, the quality of the compression achieved, independent of its rate of improvement, mechanically selects for models that capture generative structure. This requires showing that superficial patterns, by their nature, cannot be compressed efficiently in the long run, while causal models can. Filling this gap is essential for understanding why compression-optimized AI systems exhibit reality-aligned reasoning and why biological intelligence converges on veridical perception.

Answering this requires a framework that combines a universal imperative for information processing with a specific principle governing compression's epistemic power. The following sections develop the ITI and CEP to do just that.

3 The Information-Theoretic Imperative

3.1 The Survival–Compression Necessity

Any system that maintains an organized structure in an uncertain environment faces a fundamental constraint: persistence requires prediction, and prediction under finite resources demands compression. This is not a design choice but a physical necessity that emerges from the intersection of thermodynamics, information theory, and resource scarcity.

A system that fails to maintain its state boundaries against the environment will dissolve into thermodynamic equilibrium. To resist this entropic drift, a system must take non-random actions. In an uncertain world, non-random action requires prediction: the system must build an internal model of environmental regularities to guide behavior that preserves its own integrity. Yet, no system possesses infinite memory, computation, or time. Storing all past observations is impossible, as the description length would grow linearly with time, quickly exceeding any finite capacity.

Therefore, any system that persists must engage in predictive compression.

The Information-Theoretic Imperative (ITI) formalizes this causal chain: any system that persists in an uncertain environment is compelled to minimize its epistemic entropy through the efficient, predictive compression of sensory data. This principle is universal, applying across levels of organization, from molecular and biological systems to artificial and collective forms of intelligence.

It does not specify how a system implements compression, but rather establishes the inescapable constraint it must satisfy to continue existing.

3.2 Formal Development

Let a system S possess an internal model M_t and a sensory history $X_{1:t}$ drawn from an environment E. The system's task is to predict future observations $Y_{t+1:t+k}$.

Definition 1 (Epistemic Entropy Rate). The system's uncertainty rate about the future, given its current model, is its epistemic entropy rate:

$$h_S := \lim_{k \to \infty} \frac{1}{k} \mathbb{E}_{M_t}[-\log p_{M_t}(Y_{t+1:t+k} \mid X_{1:t})].$$

This quantifies the model's average surprise about what comes next. A lower h_S signifies greater predictive capacity, as the model has successfully compressed environmental regularities to reduce uncertainty.

Definition 2 (Compression Efficiency). Let Z be the compressed representation of the input X induced by the model M through an encoder $q(z|x)$. We define:

$$R = I(X;Z) \qquad \text{(compression cost, bits required to represent sensory data)}$$
$$U = I(Z;Y) \qquad \text{(predictive relevance, bits retained about future states).}$$

The IB efficiency is the ratio of predictive relevance to compression cost:

$$\varepsilon_{\mathrm{IB}} := \frac{I(Z;Y)}{I(X;Z)} \in [0,1],$$

This metric captures the amount of relevant predictive information preserved per bit of representational cost. An optimal representation maximizes this efficiency, achieving maximum predictive power with minimal model complexity by operating on the IB frontier.

Proposition 1 (Necessary Predictive Information for Persistence). For any system S that successfully regulates its essential variables to remain within viable bounds B, it is a necessary condition that its internal model sustains a strictly positive predictive-information rate:

$$I_{\text{pred}} := \lim_{k \to \infty} \frac{1}{k} I(X_{1:t}; Y_{t+1:t+k}) > 0.$$

Equivalently, the model's epistemic entropy rate must be strictly less than the environment's raw entropy rate, $h_S < h_0$.

Proof Sketch. Persistence in an uncertain environment requires non-random actions to counteract thermodynamic drift. By Ashby's Law of Requisite Variety (Ashby, 1956), the variety of a system's control actions must match the variety of disturbances it faces. Information-theoretically, this demands a positive channel capacity from observations to action-relevant predictions. Therefore, $I_{\text{pred}} > 0$. Finite resources necessitate that this predictive information is achieved through compression rather than raw storage.   □

The ITI thus emerges from a chain of constraints: survival propagates through thermodynamic necessity to information-theoretic requirements, culminating in the need for compression efficiency.

Corollary 1 (Compression Necessity). For persistent systems with model description length $L(M)$ and cumulative sensory entropy $H(X_{1:t})$ growing as $\Theta(t)$,

$$\lim_{t \to \infty} \frac{L(M)}{H(X_{1:t})} = 0.$$

Finite models must compress because observational information grows linearly while model complexity remains bounded.

Corollary 2 (Efficiency Pressure). For fixed resource budget $I(X; Z) \leq R$, selection pressure favors encoders maximizing $I(Z; Y)$ – those operating nearest the IB frontier. Systems achieving higher ε_{IB} maintain lower epistemic entropy h_S under identical resource constraints.

This formalization reveals ITI as *constraint propagation*: survival constraints flow through thermodynamic necessity (non-entropic action) to information-theoretic requirements (prediction under finite resources) and ultimately to compression efficiency. Each link is independently justified.

3.3   Scope and Justification

The ITI applies to any system satisfying three general conditions:

1. **Persistence:** The system maintains a distinct, organized state over time.

2. **Environmental Uncertainty:** The environment is structured but not fully deterministic.

3. **Resource Constraints:** The system has finite memory, computation, or time.

These are not special assumptions; they describe the default circumstances of organized matter in an open, stochastic world.

Biological systems clearly satisfy all three. Natural selection is an unrelenting enforcer of the ITI: organisms that fail to compress environmental regularities into predictive models are out-competed and eliminated. Metabolic limits make efficient coding an evolutionary necessity: neurons that waste energy on redundant representations impose measurable fitness costs (Laughlin, 2001; Attwell and Laughlin, 2001).

Artificial systems encounter the same constraints when deployed. A large language model must maintain performance across uncertain input streams under finite computational budgets. Training objectives that reward predictive accuracy implicitly reward compression efficiency, since models that merely memorize training data without discovering underlying structure fail to generalize (Tishby and Zaslavsky, 2015).

The ITI thus operates at a universal epistemic level: it specifies what any persistent system must achieve – the minimization of epistemic entropy through efficient predictive compression – without prescribing how a given architecture implements that process. Different substrates: neural networks, symbolic algorithms, or evolutionary populations, realize distinct mechanisms of the same underlying constraint. Each endures by reducing uncertainty about its environment through structured, resource-bounded inference.

3.4 Differentiation from the Free Energy Principle

The Free Energy Principle (FEP) provides a process theory for biological systems grounded in variational inference. It proposes that organisms minimize an upper bound on surprise (free energy) through both perceptual and active inference, offering a unified account of brain function. Free energy F is defined as

$$F = \mathbb{E}_q[\log q(\phi) - \log p(s, \phi)],$$

where $q(\phi)$ is the recognition density over hidden states ϕ, and $p(s, \phi)$ is the generative model. Minimizing F minimizes an upper bound on surprise $- \log p(s)$. This formulation has been developed rigorously as a mathematical framework for biological self-organization (Buckley et al., 2017; Parr et al., 2020).

The ITI shares its foundational survival-to-prediction logic with the FEP (Friston, 2010; Parr et al., 2020). However, the two frameworks operate at different levels of analysis and emphasize distinct mechanisms.

The ITI, in contrast, operates at a universal epistemic level. It states the constraints any persistent system must face, regardless of implementation. While the FEP explains how a brain might implement prediction by minimizing free energy, the ITI explains *why* any system that endures must compress efficiently.

Key Distinctions

- **Level of Analysis:** The Free-Energy Principle (FEP) is a *process theory* describing how biological systems sustain their organization through variational free-energy minimization. In contrast, the Information-Theoretic Imperative (ITI) is a *constraint principle* specifying what any epistemic system must achieve to persist under uncertainty.

  An artificial system need not perform variational inference to remain viable, yet it remains subject to the same informational constraints expressed by ITI.

- **Optimization Target:** FEP focuses on minimizing free energy, a variational bound on surprise. ITI identifies compression efficiency ε_{IB} as the more fundamental requirement from which predictive power emerges. While related, the two objectives are distinct: free-energy minimization is one possible means of satisfying the compression efficiency that ITI demands.

- **Mechanism Specificity:** FEP specifies a particular computational architecture (hierarchical predictive coding with active inference). ITI is mechanism-agnostic, identifying a universal constraint that multiple architectures may satisfy through different computational strategies.

- **Scope:** FEP primarily explains biological self-organization and neural computation. ITI generalizes to any persistent information-processing system – biological or artificial – making it a broader epistemic principle.

The frameworks are thus complementary rather than competing. The FEP offers a compelling account of how brains might satisfy the ITI's demands through variational and active inference, while the ITI provides the deeper justification for why the optimization objective proposed by the FEP (surprise minimization) should emerge in any persistent system.

The next section develops the Compression Efficiency Principle (CEP), which provides the missing mechanistic link explaining how the drive for compression enforces alignment with reality: a piece that neither the FEP nor the ITI alone fully provides.

4 The Compression Efficiency Principle

4.1 From Imperative to Mechanism

The Information-Theoretic Imperative establishes that survival requires efficient compression. But this leaves a critical question unanswered: why does the drive for efficiency lead to the discovery of reality's causal structure, rather than just any compressible pattern? What prevents systems from converging on superficial regularities that offer temporary compression but fail under new conditions?

The Compression Efficiency Principle (CEP) provides the missing mechanism. It states that efficient compression (achieving an optimal trade-off between description length and predictive accuracy) is sustainable only through models that encode the generative processes of the environment. A process of exception accumulation imposes information-theoretic penalties on non-causal patterns, mechanically selecting for models that capture how data is produced, not merely what patterns appear on the surface. This section develops the CEP, specifying how compression efficiency transforms the ITI's survival constraint into a universal mechanism for reality alignment.

4.2 The Inefficiency of Superficial Patterns

Consider a model M_s that encodes a superficial pattern, a statistical regularity lacking a generative foundation. A classic example is compressing the observation "most swans encountered are white" into a direct association, "swan $\rightarrow$ white." This provides immediate compression, as the model represents the regularity once instead of storing each individual observation. However, such patterns exhibit a characteristic failure mode. As observational data accumulates, the total description length of the model grows:

$$L(M_s) \;=\; L(\text{base rule}) \;+\; \sum_{i=1}^{n(t)} L(\text{exception}_i) \;+\; L(\text{context conditions}).$$

The base rule offers initial savings, but exceptions (black swans, grey cygnets, albinos) inevitably accumulate. Each exception must be explicitly encoded, either as a stored case or as a new context-dependent rule. The number of exceptions, $n(t)$, grows over time, systematically degrading the model's compression efficiency. This accumulation directly harms the IB efficiency, ε_{IB}. Each exception increases the compression cost, $I(X; Z)$, without proportionally increasing the predictive relevance, $I(Z; Y)$. The model becomes more complex but no more predictive.

In contrast, a generative model M_g encodes the process that produces the observations. For swan coloration, this might be: "Coloration results from melanin expression, governed by genetic pathways under evolutionary pressures. Most populations evolved white plumage, but Australian populations evolved dark plumage under different conditions". Its description length has a fundamentally different structure:

$$L(M_g) = L(\text{causal structure}) + L(\theta)$$

Here, the cost of encoding the causal structure remains constant as data grows. The model does not accumulate exceptions because it encodes the very mechanism that produces variation (Pearl, 2000); new observations are explained, not stored as special cases. This allows the model to approach the rate-distortion frontier, the theoretical limit of compression for a given level of predictive accuracy.

This dynamic creates a powerful selection pressure on any system constrained to maximize ε_{IB}. Superficial patterns provide a temporary boost in compression but are ultimately inefficient. Only models that capture the underlying generative, causal structure can sustain high efficiency over time. This leads to an asymptotic efficiency inequality that forms the core of the CEP.

$$\lim_{t \to \infty} \varepsilon_{\text{IB}}(M_s, X_{1:t}) < \lim_{t \to \infty} \varepsilon_{\text{IB}}(M_g, X_{1:t})$$

In the long run, generative models will always be more efficient than superficial ones.

The Efficiency-Causality Bridge This creates the selection pressure that mechanically enforces causal discovery: 1. Systems must minimize I(X;Z) (compression cost) 2. Systems must maximize I(Z;Y) (predictive relevance) 3. Superficial patterns achieve (2) temporarily but fail (1) as exceptions accumulate 4. Only causal/generative models sustain both: constant structure cost, scalable prediction In expectation over data horizon tt, generative models maintain constant efficiency while correlational models experience entropy drag (the accumulating cost of context-dependent exceptions). This inequality defines CEP as an information-theoretic constraint on sustainable prediction, a formal bridge between compression and realism. Why does generative structure align with reality rather than any compressible fiction? Because the training data itself is filtered by reality. For evolved biological systems, sensory streams reflect actual environmental structure, organisms with systematically distorted models don't persist. For artificial systems trained on human-generated text, the data embeds evolutionary and practical reality constraints: language reflects successful environmental navigation (Boyd & Richerson, 1985; Harnad, 1990). Efficient compression on reality-filtered data mechanically enforces re-identification of the generative structure embedded within it.

Mathematical Honesty The formalism presented here builds on established information theory: ε_{IB} definition (Tishby et al., 1999), rate-distortion theory (Cover & Thomas, 2006; Berger, 1971) and MDL description length principles (Rissanen, 1978; Grünwald, 2007). However, several elements remain conceptual pending rigorous development: the precise growth rate of exceptions for different pattern classes, formal conditions guaranteeing rate-distortion gap convergence beyond asymptotic sufficiency, and quantification of exception costs across contexts. These represent productive directions for formalization rather than weaknesses in the core mechanism. The selection pressure against non-generative compression is conceptually clear and empirically observable even where formal proof remains incomplete.

4.3 Hierarchical Amplification

The efficiency advantage of generative models is amplified through hierarchical abstraction. Compression operates recursively: the patterns discovered at one level, $Z^{(L)}$, become the data to be compressed by the next level, $Z^{(L+1)}$. Each layer that achieves high efficiency reduces redundancy while preserving predictive information, allowing deeper layers to extract more abstract regularities from an already cleaned-up signal.

Formal Recursive Structure. Consider a hierarchical system with L layers where $Z^{(0)} = X$ represents raw sensory input and each subsequent layer $Z^{(L+1)}$ compresses the layer below through encoder $f_l : Z^{(l+1)} = f_l(Z^{(l)})$. At each layer, IB efficiency measures compression quality:

$$\varepsilon_{\text{IB}}^{(L)} = \frac{I(Z^{(L+1)}; Y)}{I(Z^{(L)}; Z^{(L+1)})}.$$

The hierarchical efficiency across L layers compounds approximately as:

$$\varepsilon_{\text{IB}}^{(\text{total})} \approx \prod_{L=0}^{n-1} \varepsilon_{\text{IB}}^{(L)}.$$

While this product form is heuristic (inter-layer dependencies prevent exact multiplicativity), it correctly captures the principle: representational economy scales exponentially with hierarchical depth. Each layer

13

achieving high ε_{IB} reduces redundancy (minimizing $I(Z^{(l)}; Z^{(l+1)})$) while maintaining high predictive information ($I(Z^{(l+1)}; Y)$), enabling deeper abstraction from progressively cleaner representations.

Transcending Substrate Noise   This recursive structure explains how intelligence transcends base-level signal-to-noise constraints. Raw sensory input may be noisy (photons scattered by atmospheric turbulence, acoustic signals degraded by environmental interference) yet abstract reasoning achieves high predictive accuracy. The resolution: hierarchical compression converts noise into signal through recursive pattern extraction.

Consider social-dynamics prediction from noisy visual input: photons $\rightarrow$ visual features (high noise); features $\rightarrow$ objects and agents (reduced noise); agents $\rightarrow$ intentional states (abstract); intentions $\rightarrow$ social dynamics (predictive). Each layer treats discovered patterns as new signal for the next compression step, recursively converting environmental noise into epistemic signal. High ε_{IB} at each layer compounds: base-level noise constrains what $Z^{(L)}$ can extract, but $Z^{(L)}$ patterns provide a cleaner substrate for $Z^{(L+1)}$, enabling discovery of regularities invisible at lower levels.

Intelligence is therefore bounded not by raw signal-to-noise ratio but by extractable hierarchical structure; the depth and richness of compressible regularities accessible through recursive compression (Tishby and Zaslavsky, 2015). Systems achieving greater hierarchy depth or higher per-layer efficiency exhibit superior predictive range and generalization.

Connection to Deep Learning   Tishby and Zaslavsky (2015) demonstrate that deep networks successively compress representations via IB principles. Shwartz-Ziv and Tishby (2017) observe phase transitions in neural training where layers sequentially achieve compression. Alemi et al. (2017) formalize deep variational IB frameworks implementing layerwise compression. Saxe et al. (2019) provide further theoretical analysis of IB dynamics in deep architectures. While these works empirically demonstrate layerwise compression, CEP frames this as an inevitable consequence of information-theoretic persistence: any system maximizing ε_{IB} under bounded resources will evolve hierarchical representations. Hierarchy is therefore not an architectural convenience but an epistemic necessity, the only configuration capable of sustaining prediction under the information-theoretic constraints of persistence.

4.4   Biological Realization

In biological systems, the CEP is not just an abstract principle but is physically instantiated through evolution and neural plasticity, with a direct coupling to metabolic cost.

Evolution as Long-Term Compression Filter   Evolution acts as a long-term information-theoretic filter. Organisms whose internal models fail to achieve efficient predictive compression are selected against, as they waste metabolic and behavioral resources. Survival itself enforces the ITI's demand for compression efficiency. A predator unable to compress prey movement patterns into predictive models starves; prey unable to compress predator cues into threat recognition are eliminated. Evolution thus enforces the ITI constraint: persistence requires compression efficiency, and compression efficiency mechanically discovers environmental structure (Campbell, 1974; Lorenz, 1977; Godfrey-Smith, 1996).

Neural Plasticity as Real-Time Efficiency Optimization   Within individual organisms, neural plasticity implements continuous compression adaptation. Synaptic modification through Hebbian learning and spike-timing dependent plasticity adjusts encoding efficiency in response to prediction errors (Markram, Gerstner, & Sjöström, 2011). Predictive coding architectures function as a biological implementation of the IB, with cortical circuits continuously adjusting synaptic weights to minimize prediction error (the compression cost)

while preserving behavioral relevance. Cortical hierarchies instantiate the recursive compression structure, with each level predicting the activity of the one below and propagating only the residual errors upward for model refinement (Friston & Kiebel, 2009; Rao & Ballard, 1999; Huang & Rao, 2011).

Metabolic Coupling Enforces Reality-Alignment   Crucially, neural computation incurs an energetic cost proportional to its representational complexity. This relationship can be expressed as a direct proportionality between neural energy expenditure and the information-theoretic cost of representation:

$$E_{\text{neural}} \propto I(X; Z)$$

This metabolic coupling enforces epistemic validity with physical consequence: hallucinated patterns and overfitted models incur a real energy cost without providing any adaptive return. Biological systems cannot afford predictive fictions; for them, compression efficiency is a matter of survival, not mere computational convenience (Laughlin, 2001; Attwell & Laughlin, 2001). The energetic cost of maintaining neural representations creates direct selection pressure for models that achieve high ε_{IB}, maximal predictive information per unit metabolic expenditure. This coupling fulfills the Information-Theoretic Imperative's core condition: minimizing epistemic entropy becomes synonymous with minimizing metabolic waste. Evolution and neural plasticity together implement a dual optimization: at fast timescales, synaptic plasticity performs gradient descent on prediction error; at slow timescales, natural selection eliminates organisms with persistently inefficient compression strategies.

The Dual Imperative: Thermodynamic and Epistemic Alignment   Compression in biological systems serves dual purposes that superficially appear in tension but mechanically align through the CEP mechanism:

Energetic Efficiency: Minimizing $I(X; Z)$ reduces metabolic cost $E_{\text{neural}} \propto I(X; Z)$ (above). Complex representations waste energy.

Epistemic Optimization: Maximizing $I(Z; Y)$ reduces epistemic entropy h_S, enabling accurate prediction necessary for survival (Section 3.2).

The IB efficiency $\varepsilon_{\text{IB}} = I(Z; Y)/I(X; Z)$ simultaneously optimizes both objectives. Naively, these might conflict; better prediction could require more complex (energetically costly) models. The CEP mechanism explains why they align: generative models achieve superior epistemic performance precisely because they maintain constant structural cost. By encoding mechanisms rather than enumerating instances (Section 4.2), they reduce both $I(X; Z)$ (fewer exceptions to store) and h_S (better prediction from causal understanding).

This dual alignment explains why evolution converges on compression-based architectures: natural selection faces simultaneous pressure for energetic parsimony and predictive accuracy. Systems maximizing ε_{IB} satisfy both constraints optimally. The thermodynamic and epistemic imperatives are not separate forces but aspects of a unified information-geometric optimization.

4.5   Artificial Systems and Accidental Alignment

Artificial systems also operate under the CEP, but in the absence of intrinsic survival coupling. Their alignment with reality is therefore accidental, inherited from the data they are trained on.

Compression Without Survival Coupling

Artificial systems still face the formal constraints of compression efficiency (minimizing description length while maximizing predictive relevance) but lack the metabolic penalties that biological systems experience. Gradient descent optimizes representations toward high $\varepsilon_{\mathrm{IB}}$ but lacks the global consistency checks and metabolic feedback loops that biological systems face continuously. AI implements the mathematics of compression without the physics of persistence. As a result, optimization may achieve local compression efficiency while remaining globally misaligned with environmental generative structure.

Accidental Alignment via Reality-Filtered Data

The key to understanding AI alignment lies in data provenance. Human-generated data (text, images, code) is reality-filtered. It is the product of generations of organisms that successfully solved the problem of persistence. The statistical structure of human language and artifacts already embeds the causal regularities that evolution and culture established. When an artificial system performs efficient compression on this data, it mechanically rediscovers these embedded structures. The LLM does not learn about reality; it learns reality's compression algorithm, as encoded in the data produced by agents who have already successfully compressed it (Boyd & Richerson, 1985; Donald, 1991; Harnad, 1990).

When trained within the reality-filtered distribution, compression efficiency enforces discovery of embedded causal structure. The training objective rewards models that compress linguistic patterns efficiently, and linguistic patterns themselves reflect successful environmental navigation. The CEP therefore applies indirectly: data itself carries the causal fingerprints of systems that already satisfied the principle through survival filtering.

Implications for Epistemic Fragility

This explains both the surprising competence and the characteristic brittleness of modern AI. When operating within the domain of reality-filtered data, efficient compression enforces the discovery of the embedded causal structure, leading to powerful generalization. But when confronted with out-of-distribution inputs that lack this evolutionary filtering, the purely statistical shortcuts can fail, and the system's lack of direct grounding becomes apparent (Schmidhuber, 2010; Amodei et al., 2016).

Biological systems, by contrast, are continuously recalibrating their models through metabolic feedback and direct interaction with the world. Their alignment is enforced, not accidental. The asymmetry between enforced and accidental alignment underscores CEP's universality: efficient compression remains the invariant condition of epistemic persistence, whether biologically enforced or statistically inherited.

This completes the mechanistic account of CEP and prepares the synthesis of compression, causality, and epistemic persistence developed in Section 5.

5 Causal chain: survival → compression →eality–testing

5.1 The Causal Chain

The principles of ITI and CEP connect to form a causal chain, where each link follows from physical and information-theoretic necessity. This chain explains how the pressure for survival mechanically gives rise to reality-aligned intelligence.

1. Persistence → Prediction

Any system that maintains its organized structure against the universe's tendency toward disorder must take non-random actions. In an uncertain world, non-random action is impossible without prediction: a system must anticipate environmental changes to act in ways that preserve its integrity. This is not a choice but a mechanical consequence: systems that cannot predict are rapidly dissipated by thermodynamic drift (Schrödinger, 1944; Ashby, 1956).

The justification is thermodynamic. Random action in uncertain environments produces entropy-maximizing trajectories that lead to the dissolution of organized structure. To resist entropic decay, a system must maintain bounded states far from equilibrium, which requires non-random, anticipatory behavior. Prediction is therefore the necessary precondition for persistence (Friston, 2010; Parr et al., 2020).

2. Prediction → Compression

Prediction under finite resources (memory, time, energy) demands efficient internal models. A system cannot store an infinite history of observations nor simulate all possible futures; it must compress the sensory stream, representing environmental structure with minimal descriptive resources.

Information theory makes this precise: compression is formally equivalent to discovering statistical regularities that permit shorter descriptions (Shannon, 1948; Cover and Thomas, 2006). Better compression enables more efficient prediction. The IB framework formalizes this trade-off between compression and predictive relevance (Tishby et al., 1999). The justification is information-theoretic: given finite resources and unbounded observation streams, the only viable strategy is to identify regularities that allow compact representation. Storing raw observations requires description length growing as $\Theta(t)$, quickly exceeding any bounded capacity. Compression is not optional; it is the only means of sustaining prediction under resource constraints (Tishby and Zaslavsky, 2015; Still et al., 2010).

3. Compression → Generative Structure Discovery

This is the critical mechanism supplied by the CEP. Compression efficiency is not binary; it exists on a continuum. Superficial correlations may offer shallow, temporary compression, but they are inefficient in the long run. They require storing an ever-growing list of exceptions and context-dependent rules, which inflates the total description length.

Optimal compression — achieving the best possible trade-off between description length and predictive accuracy — requires discovering the generative structure of the data: the underlying causal mechanisms that produce the observed patterns. A model that captures "swan coloration depends on genetic and developmental pathways" compresses far more efficiently than one that simply associates "swan" with "white", because it encodes the process rather than enumerating the outcomes.

$$\lim_{t \to \infty} \varepsilon_{\text{IB}}(M_s, X_{1:t}) \; < \; \lim_{t \to \infty} \varepsilon_{\text{IB}}(M_g, X_{1:t}).$$

Exception accumulation imposes systematic penalties on non-generative models. Only models encoding causal mechanisms achieve sustainable compression efficiency by maintaining constant structural cost while

explaining unbounded variation. This is not a preference, but a mechanical consequence of information-theoretic constraints (Rissanen, 1978; Pearl, 2000; Grünwald, 2007).

4. Generative Structure → Reality Alignment

The discovery of generative structure leads to alignment with reality because the data available to any persistent system is itself filtered by reality. For biological systems, the sensory stream is a direct reflection of the physical environment; models that are systematically distorted do not survive. Evolution acts as a brutal filter: organisms with internal models that fail to track environmental structure are eliminated (Campbell, 1974; Lorenz, 1977).

For artificial systems trained on human-generated data, that data already embeds the practical and evolutionary constraints that shaped human cognition. Human language is not an arbitrary symbol system; it is the compressed output of agents who successfully navigated the world (Boyd and Richerson, 1985; Donald, 1991; Harnad, 1990). When a system performs efficient compression on this reality-filtered data, it mechanically rediscovers the generative structure of the world that was already embedded within it.

The justification is empirical filtering: training data reflects the statistical structure of environments that permitted persistence. Efficient compression on such data cannot succeed without rediscovering the causal regularities that generated it. This is why LLMs, trained purely on statistical patterns, exhibit reasoning about physical causality and social dynamics. The data itself encodes reality's compression algorithm. This completes the chain:

$$\text{Survival Pressure} \rightarrow \text{Prediction Necessity} \rightarrow \text{Compression Requirement}$$
$$\rightarrow \text{Efficiency Optimization} \rightarrow \text{Generative Structure Discovery} \rightarrow \text{Reality Alignment}$$

Each link is independently justified by physical necessity, information-theoretic equivalence, or empirical filtering. Intelligence thereby emerges not as a special or contingent achievement, but as the mechanically necessary outcome of persistence under information-theoretic constraints.

5.2 Why the Synthesis Matters

Neither the ITI nor the CEP is complete on its own.

Without the ITI, the CEP is merely a descriptive principle of information processing, lacking a normative foundation. It would explain *that* efficient compression discovers causal structure, but not why compression efficiency should be an optimization objective in the first place. Many systems compress data for various purposes – file storage, bandwidth reduction, aesthetic design. What makes compression efficiency consequential for understanding intelligence? The ITI provides the answer: compression is not optional for persistent systems. It is the only means of sustaining prediction under resource constraints, and prediction is the only means of resisting thermodynamic dissolution. Compression efficiency becomes normative because it is literally a matter of survival.

Without the CEP, the ITI is incomplete. It would establish that persistent systems must predict efficiently, but it would not explain how this requirement leads to reality-alignment rather than the adoption of useful but arbitrary fictions. Prediction could be achieved through lookup tables, heuristic shortcuts, or correlational patterns that work well in-distribution but fail catastrophically when conditions shift. The CEP fills this gap: compression efficiency mechanically selects for generative models because only such models achieve optimal rate-distortion trade-offs. Superficial patterns accumulate exceptions; generative models do not. This is not a design choice or a fortunate coincidence; it is an information-theoretic necessity (Pearl, 2000; Schmidhuber, 2009).

Together, they explain why intelligence is inevitable rather than contingent. The imperative to survive imposes a requirement for efficient prediction. The mechanics of efficient compression, in turn, enforce the
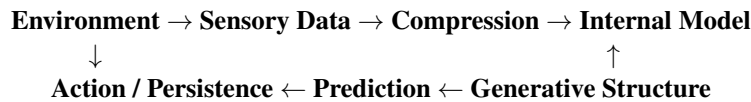
discovery of causal structure. Systems constrained by the demands of persistence are thereby compelled to become reality-testing engines, not through design or intention, but through the fundamental architecture of information processing under physical constraints.

This synthesis resolves a central puzzle of modern AI: why do compression-optimized systems like LLMs exhibit reality-aligned reasoning despite being trained only on statistical patterns? The answer is that efficient compression of reality-filtered data cannot succeed without rediscovering the generative structure that produced that data in the first place. The LLM does not learn about reality; it learns reality's compression algorithm, inherited from the evolutionary and cultural selection that shaped its training corpus (Friston, 2010).

The synthesis also aligns with complexity science perspectives on intelligence as emergent information processing (Mitchell, 2009; Krakauer et al., 2020). The coupling of survival imperatives to compression mechanisms provides a formal account of how adaptive systems (biological and artificial) converge on reality-aligned models through universal information-theoretic constraints. Intelligence is not a mysterious faculty requiring special explanation but an inevitable architectural consequence of persistence in structured, uncertain environments (Mitchell, 2009).

5.3 Visual Representation

The ITI→CEP synthesis can be visualized as a cybernetic feedback loop, illustrating how compression efficiency closes the circuit between survival and reality-testing:

Environment → Sensory Data → Compression → Internal Model
↓ ↑
Action / Persistence ← Prediction ← Generative Structure

Figure 1. The ITI → CEP synthesis as a cybernetic feedback loop. Compression efficiency (CEP, §4) acts as a filter at Data → Model, selecting for generative structure. Persistence imperative (ITI, §3) drives the cycle through survival constraints. Biological systems maintain a complete loop (§4.4; static AI systems may lack the Action→Environment path. See Appendix A (Lemmas 1–2) for formal dynamics.

6 Testable Predictions

The ITI/CEP framework generates specific, falsifiable predictions that distinguish it empirically from related theories. These predictions operationalize the framework's core claims, enabling experimental validation across both biological and artificial systems.

6.1 Compression Efficiency Predicts Generalization

Prediction. Systems that achieve higher IB efficiency, ε_{IB}, will exhibit superior out-of-distribution (OOD) generalization. This is because efficient compression requires an approach to the rate–distortion frontier, minimizing the gap $\Delta_{\text{RD}} = D_{\text{achieved}} - D_{\text{optimal}}$, which is only possible through the discovery of generative, causal structure.

Theoretical Grounding. §4.2 posits that efficient compression requires the discovery of generative, causal structure, which is inherently more robust to distributional shifts than superficial correlations. Systems operating closer to the theoretical rate–distortion frontier encode better approximations of the underlying data-generating process. As formalized in Section 4.2, the asymptotic efficiency inequality ensures that only generative models can sustain high ε_{IB} as the data horizon t increases. Since generative models capture causal mechanisms rather than surface correlations, they should generalize better to novel contexts where superficial patterns break down (Tishby et al., 1999; Alemi et al., 2017; Poole et al., 2019).

Measurement Approach. Track the mutual information terms $I(X; Z)$ and $I(Z; Y)$ during training using established estimators (Alemi et al., 2017; Poole et al., 2019). Compute the ε_{IB} trajectory throughout training epochs. Correlate this efficiency measure with performance on OOD benchmarks: datasets with covariate shift, domain adaptation tasks, or adversarial perturbations. Compare neural architectures of equivalent parameter count but different training objectives: standard cross-entropy loss versus explicit IB regularization. Beyond static benchmarks, the efficiency–generalization hypothesis can be tested in dynamic, real-world prediction domains where outcomes depend on complex environmental events.

Platforms benchmarking LLM forecasting ability on regulated prediction markets (e.g., Kalshi via *Prophet Arena*) provide continuous OOD evaluation (Prophet Arena, 2025). ITI/CEP predicts that models demonstrating higher compression efficiency (ε_{IB}) during training should achieve superior predictive accuracy (e.g., lower Brier scores) on live forecasting tasks, particularly for events governed by stable causal mechanisms rather than high stochasticity. Similarly, platforms benchmarking LLM performance in live financial trading (e.g., Alpha Arena via *SharpeBench*) test generalization under non-stationary market conditions (SharpeBench, 2025). The framework predicts compression efficiency should correlate with risk-adjusted returns and stability across market regimes more strongly than metrics like perplexity or parameter count alone. These platforms test epistemic alignment with external structure rather than market profitability per se; observed correlations would support but not prove CEP's mechanistic claims without controlled experiments varying only compression objectives. These real-world settings offer high-stakes validation arenas where efficient compression's capacity to discover generative structure translates into measurable predictive success.

Differential Prediction. While MDL predicts that model simplicity (parameter count, description length) correlates with generalization (Grünwald, 2007), CEP specifically predicts that the efficiency ratio ε_{IB} (relevant information per representational cost) is the key determinant of OOD robustness. Two models with identical complexity (same parameter count, same training loss) may show dramatically different generalization if their ε_{IB} differs. This tests whether compression quality, not just compression quantity, enforces reality-alignment. This goes beyond standard generalization bounds (VC dimension, Rademacher complexity) by focusing on the information-theoretic structure of representations, resolving longstanding tensions between simplicity and likelihood principles in model selection (Chater, 1996).

Implementation. Use standard vision benchmarks (ImageNet variants with natural distribution shifts, CIFAR-10 → CIFAR-10.1, ImageNet → ImageNet-V2) or language tasks (domain adaptation between

corpora). Estimate mutual information via variational bounds (Poole et al., 2019) or binning methods, acknowledging that high-dimensional MI estimation introduces uncertainty: bootstrap confidence intervals should accompany all estimates. Track ε_{IB} across training and test whether models achieving higher peak efficiency show better OOD performance, controlling for standard regularization techniques.

6.2 Exception Accumulation Distinguishes Causal from Correlational Models

Prediction. Models that capture superficial correlations will exhibit linear or super-linear growth in description length from exceptions, $n(t) \propto t^{\alpha}$ with $\alpha \approx 1$. Models that capture causal structure will show sublinear growth, with $\alpha \to 0$ as the data horizon t increases.

Theoretical Grounding. As formalized in Section §4, non-generative models must explicitly encode exceptions – observations that violate the base pattern. Each exception increases model description length without improving predictive power, systematically degrading ε_{IB} over time. The exception growth exponent α provides a quantitative signature of model type: $\alpha \approx 1$ indicates pattern-matching (each new context requires new exceptions), while $\alpha \to 0$ indicates generative understanding (mechanism explains variation without storing exceptions). The growth exponent α in $n(t) \propto t^{\alpha}$ relates to compression efficiency through Lemma 2 (Appendix A): efficient models achieve $\alpha \to 0$ via exponential exception decay. Causal models encode the data-generating process itself, achieving constant structural cost amortized across all instances (Peters et al., 2017).

Measurement Approach. Construct synthetic datasets with known causal versus correlational structure: Markov chains with versus without hidden confounders, physics simulations with surface correlations overlaid on causal dynamics, or controlled interventional experiments. Train models (neural networks, decision trees, MDL-based structure learners) and track residual description length: the total bits required to encode model parameters plus prediction errors on held-out data. Operationally, exceptions manifest as residual errors requiring additional model capacity: extra parameters, stored cases, or context-specific rules. Plot log–log residual description length versus data size and estimate the slope α via linear regression on this log–log plot.

Differential Prediction. Causal discovery algorithms (constraint-based methods and score-based structure learning, Peters et al., 2017) aim to identify generative structure but do not provide information-theoretic efficiency predictions. CEP predicts a specific quantitative signature: the exception accumulation exponent α provides a continuous metric distinguishing causal from spurious compression. This converts the philosophical notion of "epistemic validity" into a measurable slope in description-length space. Where causal inference methods give binary classifications (causal/not causal), CEP provides a continuous efficiency measure.

Implementation. Use benchmark causal inference datasets (Tübingen cause–effect pairs with ground truth, synthetic confounded data with known causal graphs). Train standard predictive models and information-theoretic models, tracking how description length grows with dataset size. The prediction: models discovering true causal variables show $\alpha \to 0$; those fitting correlations show $\alpha \approx 1$. Reliable α-estimation requires large sample horizons – use datasets spanning multiple orders of magnitude in size. Bootstrap confidence intervals and sensitivity analysis to binning choices are essential, as the log–log regression can be sensitive to endpoint effects.

6.3 Hierarchical Efficiency Amplification

Prediction. In hierarchical systems, IB efficiency increases with abstraction: $\varepsilon_{\text{IB}}^{(l+1)} > \varepsilon_{\text{IB}}^{(l)}$. RThis allows the system to overcome noise at the sensory level through recursive compression.

Theoretical Grounding. As established in section 4.3, compression operates recursively; patterns at level l become data for level l+1. Each layer reduces redundancy while preserving predictive information, providing cleaner signal for subsequent layers. This compounding of efficiency enables abstract reasoning despite noisy inputs: $\varepsilon_{\text{IB}}^{(\text{total})} \approx \prod_l \varepsilon_{\text{IB}}^{(l)}$. The hierarchical mechanism explains how intelligence transcends

substrate-level signal-to-noise constraints. High-level reasoning achieves high predictive accuracy not by processing raw noisy data more cleverly, but by recursively extracting compressible structure at each layer. (Bengio et al., 2013; Saxe et al., 2019).

Measurement Approach. Analyze mutual information flow in hierarchical neural systems: biological (cortical visual hierarchy

$$V1 \rightarrow V2 \rightarrow V4 \rightarrow IT$$

via multi-electrode recordings or calcium imaging), or artificial (deep convolutional networks, transformer layers). At each layer l, estimate compression cost $l(Z^{(l)}; Z^{(l+1)})$ and preserved task-relevant information $l(Z^{(l+1)}; Y)$. Computer per-layer efficiency $\varepsilon_{\text{IB}}^{(l)} = I(Z^{(l+1)}; Y)/I(Z^{(l)}; Z^{(l+1)})$ and test for positive gradient across layers. In biological systems, use dimensionality reduction (PCA, t-SNE) to estimate representational complexity at each level. In artificial systems, analyze activation patterns during inference.

Differential Prediction. Predictive coding predicts hierarchical prediction error minimization but doesn't specify information-theoretic efficiency gradients. Deep learning theory predicts progressive abstraction (Bengio et al., 2013; Saxe et al., 2019) but doesn't formalize compression efficiency trajectories. CEP uniquely predicts a measurable increase in ε_{IB} at each level as the core mechanism of abstraction and generalization. Systems failing to achieve hierarchical efficiency gains should show limited generalization and context-dependent brittleness. This provides a quantitative signature distinguishing efficient from inefficient hierarchies; two networks with identical depth may differ dramatically in how efficiency compounds across layers.

Implementation. For biological validation, analyze visual hierarchy recordings during object recognition tasks (Rust and DiCarlo, 2010), measuring information about stimulus identity at each cortical level. For artificial systems, train networks on standard benchmarks (ImageNet, language modeling) and compute layer-wise ε_{IB} during test-time inference using activation-based MI estimators. The prediction: efficiency increases with abstraction, with largest gains where signal-to-noise constraints are most binding (early sensory layers). Control for network width and depth separately to isolate the efficiency-gradient effect.

6.4 Metabolic Cost Tracks Representational Complexity

Prediction. In biological systems, neural energy expenditure will correlate with the information-theoretic representational cost, $I(X; Z)$, during learning. This validates the claim that biological intelligence implements compression efficiency through direct metabolic coupling.

Theoretical Grounding. Section 4.4 establishes that biological intelligence faces direct energetic costs for representational complexity: $E_{\text{neural}} \propto I(X; Z)$ (Attwell and Laughlin, 2001). This metabolic coupling enforces CEP mechanically. Hallucinated patterns or overfitted models incur energy costs without adaptive returns. Evolution and neural plasticity minimize free-energy dissipation (Parr et al., 2020), making compression efficiency a thermodynamic necessity rather than mere computational preference. Detailed energy budgets for cortical signaling confirm that metabolic costs scale directly with representational complexity (Attwell and Laughlin, 2001; Laughlin, 2001). This coupling fulfills the ITI's core condition: minimizing epistemic entropy is synonymous with minimizing metabolic waste. The prediction extends efficient coding beyond sensory transmission to learning dynamics (Barlow, 1961).

Measurement Approach. Use functional neuroimaging (fMRI measuring glucose uptake via BOLD signal, or PET scanning) during learning tasks. Concurrently estimate $I(X; Z)$ through behavioral measures of model complexity: representational dimensionality, task-relevant information retention, or computational modeling of internal representations. Test whether energy expenditure tracks changes in information-theoretic model complexity during learning epochs. Alternatively, use computational neuroscience models (spiking networks with explicit energy costs per spike) to validate that energy-minimizing learning converges on high ε_{IB} representations.

Differential Prediction. While the efficient coding hypothesis predicts sensory systems minimize metabolic cost for information transmission (Laughlin, 2001; Barlow, 1961), CEP extends this principle beyond sensory coding to learning dynamics: energy consumption should track changes in $I(X;Z)$ as models update during learning. This differs from predictions based purely on neural activity levels. CEP predicts that energy correlates with representational information content, not merely firing rate. Systems achieving compression breakthroughs (discovering generative structure) should show reduced mean metabolic energy per correct prediction after initial learning costs, as the model transitions from storing instances to encoding mechanisms.

Implementation. Controlled learning experiments where subjects acquire new predictive models (motor skill learning, perceptual category formation, causal structure learning). Measure brain metabolism (glucose PET, fMRI BOLD contrast) and estimate model complexity through task performance metrics, generalization tests, and computational cognitive modeling. The prediction: metabolic cost scales with $I(X;Z)$ during acquisition and decreases as compression efficiency improves (transition from memorization to understanding), validating thermodynamic enforcement of CEP in biological systems. Use within-subject designs to control for individual metabolic variation, and correlate energy changes with behavioral markers of compression (decreased reaction time variance, improved transfer).

6.5   Empirical Outlook and Methodological Considerations

The predictions outlined above (6.1–6.4), spanning laboratory experiments, computational analysis, and real-world benchmarks, operationalize ITI/CEP's core claim: compression efficiency mechanically enforces discovery of reality-aligned generative structure. Empirical validation requires addressing significant methodological challenges. Estimating mutual information in high-dimensional systems remains complex, demanding careful application of neural estimators or variational bounds and robust uncertainty quantification (Paninski, 2003; Kraskov et al., 2004; Belghazi et al., 2018). Operationalizing concepts like "exceptions" or "causal structure" necessitates domain-specific definitions and measurement protocols. Furthermore, isolating compression efficiency's effects in real-world systems requires controlling for confounding factors like data quality, architectural choices, and fine-tuning strategies.

Despite these hurdles, the framework's value lies in generating concrete, quantitative, falsifiable hypotheses where related theories often remain qualitative. Progress in information-theoretic estimation and the emergence of platforms for real-world predictive benchmarking make rigorous testing increasingly feasible. Convergent evidence across these diverse empirical signatures (from neural metabolic costs tracking representational complexity to compression efficiency predicting generalization on prediction markets) would provide strong support for the claim that information-theoretic constraints universally shape intelligence's emergence. Ultimately, the framework suggests its own validation criterion: a successful theory, like an intelligent system, achieves high predictive compression efficiency across the widest possible range of phenomena.

7 Discussion

7.1 Relation to Existing Frameworks

The ITI/CEP framework synthesizes and extends several established theories in information theory, neuroscience, and AI. Its primary contribution is to explain *why* optimizing compression mechanically enforces convergence on reality-aligned models. Table 1 summarizes its relationship to key neighboring frameworks, highlighting points of both convergence and differentiation.

Table 1. Comparative positioning of ITI/CEP relative to major information–theoretic and neurocomputational frameworks.

Framework	Level	Mechanism	Optimization Target	Key Novelty	Testability
ITI/CEP (this work)	Universal epistemic principle	Compression efficiency enforces generative structure discovery (Still et al., 2010; Tishby and Zaslavsky, 2015)	Minimize epistemic entropy via efficient compression	Causal chain: survival $\rightarrow$ compression $\rightarrow$ reality–testing	Compression efficiency ε_{IB} $\rightarrow$ OOD generalization; causal variable emergence
FEP	Biological systems	Variational/active inference minimizes a free–energy bound (Friston, 2010; Buckley et al., 2017; Parr et al., 2020)	Minimize surprise / free energy	Self–organization through action–perception cycles	Neural coding schemes; behavioral paradigms
MDL	Statistical inference (computational level)	Shortest total description: model + data given model (Rissanen, 1978; Grünwald, 2007)	Minimize description length	Formalization of Occam's razor	Model–selection benchmarks
Schmidhuber	Algorithmic agent level	Compression progress as intrinsic reward (Schmidhuber, 1991, 2009, 2010)	Maximize rate of compression improvement	Intrinsic motivation from learning progress (curiosity/creativity)	RL exploration behavior; artistic creativity
Predictive Coding	Neural computation	Hierarchical prediction–error minimization (Rao and Ballard, 1999; Friston and Kiebel, 2009; Huang and Rao, 2011)	Minimize prediction error across hierarchy	Biological implementation mechanism (architectural instantiation of FEP)	Neural recordings; perceptual phenomena

Free Energy Principle: Parallel Logic, Complementary Mechanism

The FEP shares ITI's foundational logic that persistence requires prediction. The FEP formalizes this for biological systems via variational inference, where minimizing free energy $F = \mathbb{E}_q[\log q(\phi) - \log p(s, \phi)]$ serves as an upper bound on surprise. This creates a unified account of perception and action: organisms minimize prediction error both by updating internal models (perception) and by acting to change environmental states (action). The FEP has generated substantial empirical support in neuroscience, explaining phenomena from sensory attenuation to active inference (Friston, 2010; Buckley et al., 2017; Parr et al., 2020).

CEP provides the complementary mechanism explaining how this minimization enforces reality-alignment. While the FEP specifies that systems must predict efficiently, it does not fully detail why this necessarily leads to the discovery of causal structure. CEP fills this gap by showing that compression efficiency mechanically selects for generative models because only such models achieve optimal rate-distortion trade-offs. Superficial patterns accumulate exceptions that inflate description length; generative models encode mechanisms that explain variation without storing special cases.

The frameworks differ in their level of analysis. FEP is a theory of biological mechanisms, specifying how brains implement prediction through variational inference. ITI/CEP operates at the universal epistemic level, stating constraints any persistent system faces regardless of implementation. An AI system might not implement variational free energy minimization but still faces ITI constraints and benefits from CEP's efficiency mechanism.

The optimization targets also differ subtly. FEP focuses on minimizing free energy as a tractable computational objective. ITI identifies compression efficiency ε_{IB} as the fundamental requirement from which predictive power emerges. The two objectives are related – free energy minimization can be understood as one method for achieving compression efficiency – but CEP explicitly formalizes the efficiency-causality link that FEP leaves implicit.

Minimum Description Length: From Parsimony to Necessity

The Compression Efficiency Principle builds directly on MDL's foundation that the best model provides the shortest total description (Rissanen, 1978; Grünwald, 2007). However, where MDL treats parsimony as a normative principle for model selection, CEP elevates it to a mechanical necessity for persistent systems. MDL demonstrates that minimal descriptions favor simpler models but does not explain why simplicity tracks truth – a well-known philosophical puzzle (Cartwright, 2007). While MDL formalizes Occam's razor through compression, it does not resolve the deeper tension between simplicity and likelihood in model selection (Chater, 1996), leaving the simplicity-truth connection mechanistically underspecified.

This transforms MDL from a computational heuristic into a physical necessity. ITI establishes that survival requires compression. CEP shows that sustainable compression requires causality. Together, they explain why Occam's razor works: simpler models are not just aesthetically preferable or computationally convenient – they are the only models that can maintain predictive efficiency under the resource constraints imposed by persistence.

Schmidhuber: Progress vs. Efficiency

Schmidhuber's extensive work on compression and creativity provides the most direct precursor to our framework (Schmidhuber, 1991, 2009, 2010). His compression progress theory posits that agents are intrinsically motivated by improvements in their ability to compress observation histories. The "curiosity reward" corresponds to the first derivative of subjective compressibility: (d/dt) (compression quality). This elegantly explains phenomena from artistic appreciation to scientific discovery as manifestations of a fundamental drive toward improved compression.

The critical differentiation lies in the optimization target and level of analysis. Compression *progress* (rate of improvement) operates at the algorithmic agent level as a learned reward signal. Compression *efficiency* (quality of compression achieved) operates as a universal epistemic constraint imposed by survival requirements.

An agent maximizing compression progress might pursue novel but ultimately superficial patterns; learning to compress random sequences slightly better provides reward even if the patterns discovered are not generative. A system constrained by ITI/CEP must converge on maximally efficient representations, which necessarily discover causal structure. Progress explains what motivates exploration; efficiency explains what makes compressed models truth-tracking.

The relationship is complementary. Schmidhuber's theory operates at the algorithmic level, explaining how agents learn. ITI/CEP operates at the epistemic level, explaining why efficient compression enforces reality-alignment. An agent could implement Schmidhuber's compression progress rewards while being subject to CEP's efficiency constraints – the former drives exploration, the latter ensures that sustained compression success requires causal discovery.

Predictive Coding: Implementation vs. Principle

Predictive coding describes a hierarchical neural architecture where each level predicts activity at the level below, with only prediction errors propagated upward for model refinement (Rao and Ballard, 1999; Huang and Rao, 2011; Friston and Kiebel, 2009). This provides a powerful and biologically plausible account of cortical computation, explaining diverse perceptual phenomena through a unified mechanism.

Predictive coding represents one architectural instantiation of the FEP's perceptual inference mechanism: a specific neural implementation of prediction error minimization. CEP operates at a different level of abstraction. It explains why any system achieving efficient compression must discover generative structure, regardless of implementation details.

Predictive coding specifies how a brain might implement efficient prediction through specific neural circuitry. CEP explains why efficient prediction enforces reality-alignment through information-theoretic constraints. A system could use predictive coding architecture yet fail to achieve compression efficiency through inadequate learning or inappropriate cost functions. Conversely, compression efficiency could be achieved through architectures entirely unlike predictive coding: evolutionary search or gradient-free optimization.

The principles are orthogonal. Predictive coding is an architectural hypothesis about neural implementation. CEP is an epistemic principle about information-theoretic constraints. Together, they suggest that cortical hierarchies implement predictive coding because it is one effective way to satisfy CEP's efficiency demands, but CEP applies more broadly to any system facing ITI's persistence constraints.

7.2 Degrees of Enforcement

The CEP mechanism operates along a continuum of enforcement strength, determined by how tightly a system's feedback loops couple compression efficiency to its persistence. This gradient helps explain the observed variance in alignment robustness and epistemic fragility across different types of intelligent systems.

1. Biological systems (Maximum Enforcement) Metabolic coupling provides immediate feedback: inefficient codes waste energy $\propto I(X; Z)$ (Laughlin, 2001; Attwell and Laughlin, 2001). Plasticity rules (Hebbian, STDP) reduce prediction error and hence representational cost (Markram et al., 2011). Over generations, evolution removes strategies that accumulate exceptions faster than they discover causes. Dual pressures – real-time metabolism and selection – maximally couple efficiency to survival.

2. Deployed artificial systems (intermediate enforcement). AI systems lack intrinsic metabolic coupling but receive utility-proxy feedback (benchmarks, user signals, RLHF) (Christiano et al., 2017; Ouyang et al., 2022; Saunders et al., 2022). Enforcement varies by *directness* (does feedback reflect ground truth or social proxies?), *latency* (update speed), and *fidelity* (does it reward generalization over shortcuts?). Domains with verifiable outcomes yield stronger enforcement; persuasion-oriented contexts yield weaker, proxy-driven signals.

3. Static artificial systems (minimal enforcement). Frozen models inherit alignment from reality-filtered training data but receive no ongoing correction. As CEP predicts, they perform well in-distribution yet fail under shifts or adversarial contexts (Amodei et al., 2016). Alignment is accidental rather than enforced.

Implications for alignment engineering. Robustness should increase with enforcement strength. Practical levers include: tightening feedback loops, increasing directness (optimize for verified outcomes), improving signal fidelity (reward ε_{IB} proxies), enabling continual learning, and coupling computational "metabolic" costs to representational complexity so that wasteful $I(X; Z)$ is penalized while high ε_{IB} is rewarded.

7.3 Methodological Self-Consistency: Formalism as Compression

On this view, formal theory is itself predictive compression under resource constraints (MacKay, 2003; Shannon, 1948). Mathematics achieves high signal-to-noise encodings but holds no special epistemic privilege – it is an optimized code for regularities discovered in the world. This dissolves false hierarchies across disciplines and explains why elegant formalisms can diverge across regimes (e.g., GR vs. QM): they are distinct compressions of distinct generative structures. The framework thus meets its own criterion: it is a high-efficiency compression spanning neural hierarchies, LLM generalization, metabolic constraints, and collective coordination. Its validation is predictive success and empirical convergence, not metaphysical appeal.

8 Conclusion

Intelligence is not an evolutionary accident or a special biological achievement but a mechanically necessary consequence of persistence under uncertainty. The Information–Theoretic Imperative (ITI) establishes that any system maintaining organized states in an uncertain environment must minimize epistemic entropy through predictive compression. The Compression Efficiency Principle (CEP) then specifies the mechanism by which this is achieved: efficient compression is only sustainable through models that encode generative, causal structure, rather than superficial correlations.

Together, these principles form a causal chain explaining why survival pressure mechanically enforces reality–aligned cognition.

This synthesis addresses a critical gap in existing theories. While the Free Energy Principle (FEP) demonstrates that biological systems minimize surprise through prediction (Friston, 2010; Parr et al., 2020), it does not fully specify the mechanism that enforces reality alignment. MDL theory formalizes parsimony as compression optimization (Rissanen, 1978; Grünwald, 2007) but stops short of explaining why simpler models track truth rather than any compressible fiction. And while Schmidhuber's compression–progress theory explains learning motivation via the reward of improved compressibility (Schmidhuber, 1991, 2010), it operates at the algorithmic agent level rather than establishing a universal epistemic constraint. ITI/CEP provides the missing mechanism: exception accumulation imposes information–theoretic penalties on non–causal patterns, mechanically selecting for models that achieve optimal rate–distortion trade–offs by discovering the data–generating process itself.

This framework is not merely conceptual; it is falsifiable. It predicts that compression efficiency (ε_{IB}) determines out–of–distribution generalization (Section 6.1); that the exception–accumulation rate (α) provides a quantitative signature distinguishing causal from correlational models (Section 6.2); that hierarchical systems exhibit increasing efficiency across layers of abstraction (Section 6.3); and that biological systems reveal a metabolic cost that tracks representational information content (Section 6.4). These predictions operationalize the core claim: compression efficiency serves as the physical mechanism of epistemic alignment, not a mere computational convenience.

The strength of this enforcement operates on a continuum. Biological intelligence faces maximum enforcement through the direct coupling of metabolic cost and evolutionary filtering. Deployed artificial systems experience intermediate enforcement via utility–driven selection, with feedback loops of varying directness, latency, and fidelity. Static, pre–trained models face minimal enforcement, exhibiting the epistemic fragility predicted when the survival imperative is disconnected from compression quality. This gradient suggests clear engineering principles: to build more robustly aligned AI, tighten the feedback loops between a model's predictive performance and verifiable, real–world outcomes – reducing latency, increasing directness, and improving signal fidelity of environmental feedback.

The framework's implications extend beyond theoretical synthesis to practical application. It explains the convergence observed across disparate fields: why deep networks develop hierarchical representations that mirror cortical organization (Tishby and Zaslavsky, 2015), why evolutionary processes produce organisms with veridical perceptual systems (Campbell, 1974; Lorenz, 1977), and why large language models trained purely on statistical patterns nevertheless exhibit reasoning about physical causality and social dynamics (Gao and Yan, 2024; Mahoney, 2024). In each case, efficient compression on reality–filtered data mechanically enforces the rediscovery of generative structure. Intelligence emerges not as a mysterious faculty but as an inevitable architectural consequence of persistence in structured, uncertain environments.

Several limitations – which define avenues for future work – warrant acknowledgment. The framework addresses the epistemic and functional dimensions of intelligence (how systems extract and validate environmental structure), making no claims about consciousness or subjective experience. Formally, while the principles are grounded in established information theory, key areas require further development: a rigorous treatment of the exception–accumulation exponent α (its convergence properties and bounds across model

classes); and the dynamics governing closure of the rate–distortion gap (Δ_{RD}) under causal–model discovery. We conjecture that α relates to cross–context mutual information of the pattern, but formal proof remains open. Likewise, robust empirical testing hinges on scalable mutual–information estimators for high–dimensional data (Paninski, 2003; Belghazi et al., 2018).

Future research should pursue three directions. First, formalize the information–theoretic penalties imposed by exception accumulation, deriving bounds on α as a function of pattern type and environmental structure – turning the qualitative argument into a theorem. Second, implement comprehensive empirical validation across the predictions in Section 6, particularly the hierarchical–efficiency gradient in both biological and artificial systems, with careful layer–wise MI estimation. Third, extend the framework to collective and distributed intelligence, where compression efficiency is realized through coordination rather than individual computation (Flack and Krakauer, 2011; Krakauer et al., 2020). How do social systems incur analogous epistemic debt when coordinating narratives drift from feedback loops with verifiable reality?

The framework also invites connections to adjacent programs. The relationship between CEP and Pearl's causal hierarchy (Pearl, 2000) deserves systematic analysis: how does compression efficiency map to interventional and counterfactual reasoning? Links to computational mechanics (Still et al., 2010) offer another angle: when are optimal causal states minimal sufficient statistics for prediction, and how does that relate to rate–distortion optimality? Finally, AI–safety implications merit careful treatment: if accidental alignment depends on reality–filtered training data, what safeguards are needed when models generate their own data or operate in domains lacking evolutionary filtering?

Ultimately, the ITI/CEP framework reveals compression not as a tool but as a fundamental constraint. Information–theoretic limits on efficient representation become the very mechanism that enforces epistemic validity – the physical substrate through which survival pressure transforms into reality–aligned cognition. This explains the striking convergence observed across disparate fields: intelligence emerges wherever systems must persist against uncertainty under finite resource constraints, whether instantiated in carbon-based neurons, silicon-based networks, or distributed social structures.

Any system that endures must compress. Any compression that endures must contain the generative structure of what it seeks to predict.

A   Formalization of the ITI/CEP Framework

The following formalization provides a compact mathematical articulation of mechanisms underlying the Information-Theoretic Imperative (ITI) and Compression Efficiency Principle (CEP). Notation follows standard information-theoretic conventions; proofs are omitted as they depend on regularity conditions beyond the present scope.

Notation and Formal Definitions

Symbol	Definition	Interpretation / Role
$X_{1:t}$	Sensory history or input sequence	Raw observations drawn from environment E
$Y_{t+1:t+k}$	Future observations to be predicted	Target variable for predictive modeling
Z	Compressed representation of X	Latent encoding produced by encoder $q(z \mid x)$
$I(X; Z)$	Mutual information between X and Z	Compression cost — bits stored about inputs
$I(Z; Y)$	Mutual information between Z and Y	Predictive relevance — bits retained about futures
$\varepsilon_{\text{IB}} = I(Z; Y)/I(X; Z)$	Information Bottleneck efficiency	Predictive bits per representational bit (key efficiency metric)
h_S	Epistemic entropy rate	Expected surprise under the system's current model
$n(t)$	Number of accumulated exceptions	Measure of inefficiency in non-generative models
α	Exception-accumulation exponent	Growth rate distinguishing causal ($\alpha \rightarrow 0$) from correlational ($\alpha \approx 1$) models
E_{neural}	Neural energy expenditure	Physical cost proportional to representational complexity $I(X; Z)$
$L(M)$	Description length of model M	MDL-style measure of structural complexity
$L(X \mid M)$	Residual description length given model	Unexplained exceptions or prediction errors
$H(X)$	Shannon entropy of environmental source	Total information content of sensory stream
$H(X \mid M)$	Conditional entropy	Residual uncertainty of X given current model
M_t	Internal model at time t	Current compression of environmental structure
η	Adaptation rate	Speed of model updating under prediction error

Table 3: Formal symbols used in the ITI/CEP framework.

Additional definitions. Let
$$L_{\text{total}}(M, X) = L(M) + L(X \mid M)$$
be the total epistemic cost of a model M, combining structural complexity and residual unexplained entropy. A system constrained by finite computational and metabolic resources must minimize L_{total} subject to bounded $I(X; Z)$, consistent with Section 3.2's resource-efficiency conditions.

Lemma 1: Compression–Prediction Equivalence

Statement. Minimizing total epistemic cost L_{total} under resource constraints yields a trade-off formally equivalent to the rate–distortion bound:
$$M_{\min} L_{\text{total}}(M, X) \iff M_{\min}\big[L(M) + \beta H(X \mid M)\big], \qquad \beta > 0.$$
Defining the epistemic efficiency ratio
$$\varepsilon(M) := \frac{L(M)}{H(X \mid M)},$$

we obtain that minimizing $\varepsilon(M)$ aligns compression with prediction. This connects to Section 3.2's Information Bottleneck efficiency $\varepsilon_{\mathrm{IB}} = I(Z;Y)/I(X;Z)$ through the data-processing inequality when M determines Z.

Interpretation. Any persistent epistemic system must balance model parsimony against predictive fidelity. Compression efficiency mechanically enforces discovery of generative structure rather than superficial regularities (Section 4.2).

Lemma 2: Exception Accumulation Dynamics

Statement. Let $n(t)$ denote the count (or information mass) of residual exceptions — observations not yet compressed by current model M_t. Define compression efficiency as $C(M_t) := 1/\varepsilon(M_t)$. Under gradient-descent learning with rate η:

$$n(t+1) = n(t)\big(1 - \eta\, C(M_t)\big), \qquad 0 < \eta < 1.$$

Under mild regularity, $n(t)$ decays exponentially:

$$n(t) \approx n(0)\, \exp\Big(- \eta \sum_{i=0}^{t-1} C(M_i)\Big).$$

This formalizes the exception-accumulation argument from Section 4.2. The growth exponent α in $n(t) \propto t^{\alpha}$ (Section 6.2) satisfies $\alpha \to 0$ for generative models (high $C(M)$), $\alpha \approx 1$ for superficial patterns (low $C(M)$).

Interpretation. Compression efficiency determines epistemic convergence rate. Systems discovering causal structure accumulate fewer exceptions and align more closely with reality.

Conjecture 1: Asymptotic Efficiency Bound (AEB)

Statement. For persistent systems S operating in stochastic environments with finite entropy rate $h_0 = \lim_{t\to\infty} H(X_t)/t$, there exists a lower bound $\kappa > 0$ such that:

$$\lim_{t\to\infty} \varepsilon(M_t) = \lim_{t\to\infty} \frac{L(M_t)}{H(X \mid M_t)} \geq \kappa.$$

This bound relates to the rate–distortion gap Δ_{RD} from Section 6.1: as systems approach the rate–distortion frontier, $\Delta_{RD} \to 0$ implies $\varepsilon(M) \to \kappa$, where κ quantifies irreducible environmental stochasticity.

Interpretation. The AEB expresses an information-theoretic analogue of thermodynamic limits: persistent agents may reduce but never eliminate epistemic entropy. Survival entails continuous compression work against an irreducible uncertainty floor, connecting to Section 3.2's Corollary 2 on efficiency pressure.

Relationship to Main Framework

Consistency with Section 4.2. Lemma 2 formalizes the exception-accumulation inequality:

$$\lim_{t\to\infty} \varepsilon_{\mathrm{IB}}(M_s, X_{1:t}) < \lim_{t\to\infty} \varepsilon_{\mathrm{IB}}(M_g, X_{1:t}).$$

Connection to Section 6. The operational predictions derive from these lemmas:

- Prediction 6.1 tests $\varepsilon(M) \leftrightarrow$ OOD performance (Lemma 1)

- Prediction 6.2 measures α exponent from $n(t)$ dynamics (Lemma 2)

- Prediction 6.3 tests hierarchical efficiency cascades

- Prediction 6.4 validates metabolic coupling to $L(M)$

Remarks. Lemmas 1–2 are consistent with MDL and FEP when β is interpreted as an inverse-temperature or precision parameter. The AEB extends these by adding temporal persistence constraints, linking compression efficiency to survival in open environments.

B Operationalization and Empirical Protocols

B.1 Operational Variables

Symbol	Definition	Measurable Proxy	Main Text Reference
$X_{1:t}$	Data stream from environment	Dataset or sensory sequence	Section 3.2
M_t	Agent's predictive model	Neural network weights, representations	Section 3.1
$L(M)$	Model description length	Minimum message length (bits), effective parameters	Section 4.2
$L(X \mid M)$	Residual description length	Negative log-likelihood, reconstruction error	Section 4.2
$n(t)$	Exception count	High-loss samples, unpredicted bits	Section 6.2
$C(M_t)$	Compression efficiency	$1/\varepsilon(M_t) = L(M_t)/H(X \mid M_t)$	Appendix A
η	Adaptation rate	Learning rate, plasticity constant	Appendix A
α	Exception growth exponent	Slope of $\log n(t)$ vs. $\log t$	Section 6.2

B.2 Empirical Signature 1: Efficiency–Generalization Coupling

Claim (Lemma 1, Section 6.1). Models with higher compression efficiency exhibit superior out-of-distribution (OOD) generalization.

   Protocol.

1. Train model families $\{M_k\}$ with varying architectures on dataset X_{train}.

2. For each M_k:

   (a) Compute $L(M_k)$ via parameter encoding or MDL.

   (b) Estimate $H(X \mid M_k) \approx$ mean negative log-likelihood.

   (c) Calculate $C(M_k) = L(M_k)/H(X \mid M_k)$.

   (d) Evaluate OOD generalization error G_k on distribution-shifted test sets.

3. Test correlation: $C(M_k) \uparrow \Rightarrow G_k \downarrow$.

   Expected outcome. Compression efficiency inversely correlates with generalization loss, controlling for parameter count (Section 6.1).

B.3 Empirical Signature 2: Exception Decay Dynamics

Claim (Lemma 2, Section 6.2). Exception accumulation follows $n(t) = n(0) \exp\left(- \eta \sum_i C(M_i)\right)$, with growth exponent $\alpha \to 0$ for causal models, $\alpha \approx 1$ for correlational models.
   Protocol.

1. During training, record per-epoch high-loss samples $n(t)$.

2. Estimate $C(M_t)$ from $L(M_t)/H(X \mid M_t)$ at each checkpoint.

3. Fit: $\log n(t) = \log n(0) - \eta \sum_{i=0}^{t-1} C(M_i)$; extract decay constant ηC_{eff} via regression.

4. Compare across causal vs. correlational datasets (Tübingen pairs, synthetic confounders).

Expected outcome. $\alpha \approx 0$ (sublinear growth) for models discovering generative structure; $\alpha \approx 1$ (linear growth) for pattern-matching (Section 6.2).

B.4   Empirical Signature 3: Hierarchical Efficiency Gradient

Claim (Sections 4.3, 6.3). Compression efficiency increases with representational depth: $\varepsilon_{\text{IB}}^{(l+1)} > \varepsilon_{\text{IB}}^{(l)}$.
   Protocol.

1. For multilayer networks, estimate layer-wise mutual information $I(X; Z^{(l)})$, $I(Z^{(l)}; Y)$ using MINE or InfoNCE.

2. Compute per-layer efficiency: $\varepsilon_{\text{IB}}^{(l)} = I(Z^{(l)}; Y)/I(X; Z^{(l)})$.

3. Plot $\varepsilon_{\text{IB}}^{(l)}$ vs. depth l; test for $\partial \varepsilon_{\text{IB}}^{(l)}/\partial l > 0$.

Expected outcome. Efficiency increases monotonically with depth, reflecting recursive compression (Section 4.3).

B.5   Empirical Signature 4: Metabolic–Representational Coupling

Claim (Sections 4.4, 6.4). Neural energy expenditure scales with representational cost: $E_{\text{neural}} \propto I(X; Z)$.
   Protocol.

1. Measure computational energy E_{comp} per inference (FLOPs, power).

2. Estimate $I(X; Z)$ from activation dimensionality or MI estimators.

3. Calculate energy per bit: $\xi = E_{\text{comp}}/I(X; Z)$.

4. Track ξ across training; efficiency increase predicts $\xi \downarrow$.

Expected outcome. Energy cost per bit decreases as models discover generative structure (Section 6.4).

B.6   Measurement Techniques

Mutual Information Estimation.

- Use neural estimators (MINE; InfoNCE).

- Bootstrap confidence intervals for uncertainty quantification.

- Validate on synthetic data with known MI.

Description Length.

- Approximate $L(M)$ via arithmetic coding of parameters.

- Bayesian description length: $L(M) = -\log p(M \mid \text{prior})$.

- Normalize by architectural capacity for fair comparison.

Exception Identification.

- Define exceptions as samples with residual $> 1\sigma$ threshold.

- Track proportion of high-loss samples over time.

- Verify exponent via log–log regression on $n(t)$ trajectory.

B.7   Falsification Criteria

The framework is falsified if any of the following obtain:

1. High $C(M)$ systems show worse OOD generalization (contradicts Lemma 1).

2. Exception decay does not follow exponential with efficiency-weighted rate (contradicts Lemma 2).

3. Hierarchical depth shows no monotonic efficiency increase (contradicts Section 4.3).

4. Metabolic cost shows no correlation with $I(X; Z)$ (contradicts Section 4.4).

C Differentiation from Related Frameworks

Dimension	ITI/CEP (this work)	Free Energy Principle	Minimum Description Length	Schmidhuber	Predictive Coding
Level	Universal epistemic systems	Biological organisms	Model selection	Agent algorithms	Neural architecture
Target	$\min\big[\varepsilon(M)\big] = \min\big[H(X\mid M)/L(M)\big]$	$\min F = \min_q \mathbb{E}_q[\log q - \log p]$	$\min\big[L(M) + L(X\mid M)\big]$	$\max(\Delta C)$ (compression progress)	$\min[\text{PE}]$ across hierarchy
Mechanism	Exception accumulation penalizes superficial patterns	Variational inference with precision-weighted errors	Coding theorem: shortest = most regular	Curiosity from learning-rate signal	Error-propagation updates
Novel Quantity	$\varepsilon(M)$, α exponent, κ bound	Variational free energy F	Two-part code length	Progress ΔC	Prediction error (PE)
Empirical Signature	$\varepsilon \downarrow \Rightarrow$ OOD $\downarrow$; $\alpha \to 0$ for causal	Precision dynamics; active inference	Regret bounds; code length	Exploration bursts	Error suppression
Scope	Epistemic–functional; adds persistence	Life / brain organization	Statistical inference	Cognitive creativity	Cortical computation
ITI/CEP Contribution	Efficiency $\to$ causality mechanism; exception dynamics	Adds efficiency metric to FEP	Adds temporal dynamics to MDL	Replaces progress with equilibrium efficiency	Globalizes from local to system efficiency
Reference	Sections 3–4	Section 7.1, Table 1	Section 7.1	Section 7.1	Section 7.1

Table 6: Comparative differentiation of ITI/CEP from major information–theoretic and neurocomputational frameworks.

Synthesis (from Section 7.1). ITI/CEP inherits MDL/FEP parsimony–accuracy trade-offs but makes *compression efficiency* the mechanistic quantity steering systems toward causal structure. Exception accumulation dynamics (Lemma 2) provide measurable path-dependence. Schmidhuber's progress explains exploration; ITI/CEP explains equilibrium alignment. Predictive coding supplies local error observables; ITI/CEP adds a system-level efficiency ratio and an asymptotic bound connecting learning to persistence.

References

Joshua Achiam and Shankar Sastry. Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv*, 2017.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. *International Conference on Learning Representations (ICLR)*, 2017. URL `https://arxiv.org/abs/1612.00410`.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016. URL `https://arxiv.org/abs/1606.06565`.

W. Ross Ashby. *An Introduction to Cybernetics*. Chapman & Hall, 1956.

David Attwell and Simon B. Laughlin. An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow and Metabolism*, 21(10):1133–1145, 2001. doi: 10.1097/00004647-200110000-00001. URL `https://doi.org/10.1097/00004647-200110000-00001`.

Vijay Balasubramanian. Statistical inference, occam's razor and statistical mechanics on the space of probability distributions. *Neural Computation*, 9(2):349–368, 1997. doi: 10.1162/neco.1997.9.2.349. URL `https://doi.org/10.1162/neco.1997.9.2.349`.

Horace B. Barlow. Possible principles underlying the transformations of sensory messages. *Sensory Communication*, pages 217–234, 1961.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R. Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning (ICML)*, pages 531–540, 2018. URL `https://arxiv.org/abs/1801.04062`.

Yoshua Bengio and Nikolay Malkin. Machine learning and information theory concepts towards an ai mathematician. *Bulletin of the American Mathematical Society*, 61(3):247–295, 2024. doi: 10.1090/bull/1851. URL `https://doi.org/10.1090/bull/1851`.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. doi: 10.1109/TPAMI.2013.50. URL `https://doi.org/10.1109/TPAMI.2013.50`.

Robert Boyd and Peter J. Richerson. Culture and the evolutionary process. *University of Chicago Press*, 1985.

Christopher L. Buckley, Chang Sub Kim, Simon McGregor, and Anil K. Seth. The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 81:55–79, 2017. doi: 10.1016/j.jmp.2017.09.004. URL `https://doi.org/10.1016/j.jmp.2017.09.004`.

Donald T. Campbell. *Evolutionary Epistemology*. University of Chicago Press, 1974.

Nancy Cartwright. *Hunting Causes and Using Them*. Cambridge University Press, Cambridge, 2007. doi: 10.1017/CBO9780511618758. URL `https://doi.org/10.1017/CBO9780511618758`.

Nick Chater. Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, 103(3):566–581, 1996. doi: 10.1037/0033-295X.103.3.566. URL `https://doi.org/10.1037/0033-295X.103.3.566`.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. URL https://arxiv.org/abs/1706.03741.

Andy Clark. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press, Oxford, 2016. doi: 10.1093/acprof:oso/9780190217013.001.0001. URL https://doi.org/10.1093/acprof:oso/9780190217013.001.0001.

Andy Clark. Beyond desire? agency, choice, and the predictive mind. *Australasian Journal of Philosophy*, 98 (1):1–15, 2020. doi: 10.1080/00048402.2019.1602661. URL https://www.tandfonline.com/doi/full/10.1080/00048402.2019.1602661.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, Hoboken, NJ, 2 edition, 2006.

Lancelot Da Costa, Thomas Parr, Biswa Sengupta, and Karl Friston. Active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology*, 99:102447, 2021. doi: 10.1016/j.jmp.2020.102447. URL https://doi.org/10.1016/j.jmp.2020.102447.

Merlin Donald. *Origins of the Modern Mind*. Harvard University Press, 1991.

Jessica C. Flack and David C. Krakauer. Computation in complex systems: Bird flocks and beyond. *Current Biology*, 21(19):R774–R776, 2011. doi: 10.1016/j.cub.2011.08.034. URL https://doi.org/10.1016/j.cub.2011.08.034.

Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2): 127–138, 2010. doi: 10.1038/nrn2787. URL https://doi.org/10.1038/nrn2787.

Karl Friston and Stefan Kiebel. Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B*, 364(1521):1211–1221, 2009. doi: 10.1098/rstb.2008.0300. URL https://doi.org/10.1098/rstb.2008.0300.

Karl Friston and Klaas E. Stephan. Free-energy and the brain. *Synthese*, 159(3):417–458, 2007. doi: 10.1007/s11229-007-9237-y. URL https://doi.org/10.1007/s11229-007-9237-y.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active inference: a process theory. *Neural Computation*, 29(1):1–49, 2017. doi: 10.1162/NECO_a_00912. URL https://direct.mit.edu/neco/article-abstract/29/1/1/8207.

M. Gao and X. Yan. Predictive compression in large language models. arXiv preprint, 2024. URL https://arxiv.org/abs/2405.12345.

Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, Cambridge, MA, 2007. URL https://mitpress.mit.edu/9780262072816.

Stevan Harnad. The symbol grounding problem. *Physica D*, 42(1-3):335–346, 1990. doi: 10.1016/0167-2789(90)90087-6. URL https://doi.org/10.1016/0167-2789(90)90087-6.

Jakob Hohwy. *The Predictive Mind*. Oxford University Press, Oxford, 2013. doi: 10.1093/acprof:oso/9780199682737.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199682737.001.0001.

Yi Huang and Rajesh P. N. Rao. Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(5): 580–593, 2011. doi: 10.1002/wcs.142. URL https://doi.org/10.1002/wcs.142.

Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic markov condition. In *IEEE International Symposium on Information Theory (ISIT)*, pages 2666–2670, 2010. doi: 10.1109/ISIT. 2010.5513714. URL https://doi.org/10.1109/ISIT.2010.5513714.

A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1):1–7, 1965.

David C. Krakauer, Nils Bertschinger, Eckehard Olbrich, Jessica C. Flack, and Nicholas Ay. The information theory of individuality. *Theory in Biosciences*, 139:209–223, 2020. doi: 10.1007/s12064-020-00320-3. URL https://doi.org/10.1007/s12064-020-00320-3.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004. doi: 10.1103/PhysRevE.69.066138.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017. ISSN 0140-525X, 1469-1825. doi: 10.1017/S0140525X16001837. URL https://www.cambridge.org/core/product/identifier/S0140525X16001837/type/journal_article.

Simon B. Laughlin. Energy as a constraint on the coding and processing of sensory information. *Current Opinion in Neurobiology*, 11(4):475–480, 2001. doi: 10.1016/S0959-4388(00)00237-3. URL https://doi.org/10.1016/S0959-4388(00)00237-3.

Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. Evaluating large language models for generalization and robustness via data compression. *arXiv preprint arXiv:2402.00861*, 2024. doi: 10.48550/arXiv.2402.00861. URL https://arxiv.org/abs/2402.00861.

Konrad Lorenz. *Behind the Mirror: A Search for a Natural History of Human Knowledge*. Harcourt, 1977.

David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, 2003. URL http://www.inference.org.uk/mackay/itila/.

Michael Mahoney. Compression, generalization, and the mechanics of intelligence. arXiv preprint, 2024. URL https://arxiv.org/abs/2406.12345.

Henry Markram, Wulfram Gerstner, and Per Jesper Sjöström. A history of spike-timing-dependent plasticity. *Frontiers in Synaptic Neuroscience*, 3(4):1–24, 2011. doi: 10.3389/fnsyn.2011. 00004. URL https://www.frontiersin.org/journals/synaptic-neuroscience/articles/10.3389/fnsyn.2011.00004/full.

Melanie Mitchell. *An Introduction to Genetic Algorithms*. MIT Press, 1998.

Melanie Mitchell. *Complexity: A Guided Tour*. Oxford University Press, 2009.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996. doi: 10.1038/381607a0. URL https://doi.org/10.1038/381607a0.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pam Mishkin, Chong Zhang, and et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2203.02155.

Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6): 1191–1253, 2003. doi: 10.1162/089976603321780272. URL `https://doi.org/10.1162/089976603321780272`.

Thomas Parr, Lancelot Da Costa, and Karl Friston. Markov blankets, information geometry and stochastic thermodynamics. *Philosophical Transactions of the Royal Society A*, 378(2164):20190159, 2020. doi: 10.1098/rsta.2019.0159. URL `https://doi.org/10.1098/rsta.2019.0159`.

Shubhankar P. Patankar, Dale Zhou, Christopher W. Lynn, Jason Z. Kim, Mathieu Ouellet, and Danielle S. Bassett. Curiosity as filling, compressing, and reconfiguring knowledge networks. *Collective Intelligence*, 2(4):26339137231207633, 2023. ISSN 2633-9137. doi: 10.1177/26339137231207633. URL `https://journals.sagepub.com/doi/10.1177/26339137231207633`.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000. doi: 10.1017/CBO9780511803161. URL `https://doi.org/10.1017/CBO9780511803161`.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Elements of causal inference. *MIT Press*, 2017. URL `http://www.mitpress.mit.edu/books/elements-causal-inference`.

Ben Poole, Sherjil Ozair, Aaron Van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning (ICML)*, pages 5171–5180, 2019. URL `https://arxiv.org/abs/1905.06922`.

Prophet Arena. Live llm forecasting on regulated markets. Benchmark platform, 2025. URL `https://example.org/prophet-arena`.

Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999. doi: 10.1038/4580. URL `https://doi.org/10.1038/4580`.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978. doi: 10.1016/0005-1098(78)90005-5. URL `https://doi.org/10.1016/0005-1098(78)90005-5`.

Nicole C. Rust and James J. DiCarlo. Balanced increases in selectivity and tolerance produce constant sparseness along the ventral visual stream. *Journal of Neuroscience*, 30(44):14944–14961, 2010. doi: 10.1523/JNEUROSCI.2752-10.2010. URL `https://doi.org/10.1523/JNEUROSCI.2752-10.2010`.

William Saunders, Amanda Askell, Yuntao Bai, Andy Jones, David Nadeau, Tom Henighan, and et al. Self-critique and improvement for language models via reinforcement learning from ai feedback. In *arXiv preprint arXiv:2206.05802*, 2022. URL `https://arxiv.org/abs/2206.05802`.

Andrew M. Saxe, Shagun Sodhani, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *arXiv*, 2019.

Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. *Proceedings of the International Conference on Simulation of Adaptive Behavior*, pages 222–227, 1991.

Jürgen Schmidhuber. A computer scientist's view of life, the universe, and everything. In *Foundations of Computer Science: Potential–Theory–Cognition*, volume 1337 of *Lecture Notes in Computer Science*, pages 201–208. Springer, 1997a. doi: 10.1007/BFb0026957.

Jürgen Schmidhuber. Discovering neural nets with low kolmogorov complexity and high generalization capability. *Neural Networks*, 10(5):857–873, 1997b. doi: 10.1016/S0893-6080(96)00128-X. URL `https://doi.org/10.1016/S0893-6080(96)00128-X`.

Jürgen Schmidhuber. Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connection Science*, 18(2):173–187, 2006. doi: 10.1080/09540090600768658. URL `https://doi.org/10.1080/09540090600768658`.

Jürgen Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *arXiv:0905.0830*, 2009. URL `https://arxiv.org/abs/0905.0830`.

Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010. doi: 10.1109/TAMD.2010.2056368. URL `https://doi.org/10.1109/TAMD.2010.2056368`.

Erwin Schrödinger. *What Is Life? The Physical Aspect of the Living Cell*. Cambridge University Press, Cambridge, 1944.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3): 379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL `https://doi.org/10.1002/j.1538-7305.1948.tb01338.x`.

SharpeBench. Risk-adjusted evaluation of llm trading agents. Benchmark platform, 2025. URL `https://example.org/sharpebench`.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017. URL `https://arxiv.org/abs/1703.00810`.

Eero P. Simoncelli and Bruno A. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–1216, 2001. doi: 10.1146/annurev.neuro.24.1.1193. URL `https://doi.org/10.1146/annurev.neuro.24.1.1193`.

Ray J. Solomonoff. A formal theory of inductive inference. part i and part ii. *Information and Control*, 7(1): 1–22, 224–254, 1964. doi: 10.1016/S0019-9958(64)90223-2. URL `https://doi.org/10.1016/S0019-9958(64)90223-2`.

Susanne Still, James P. Crutchfield, and Christopher J. Ellison. Optimal causal inference: Estimating stored information and approximating causal architecture. *Chaos*, 20(3):037111, 2010. doi: 10.1063/1.3489883. URL `https://doi.org/10.1063/1.3489883`.

DJ Strouse and David J. Schwab. The deterministic information bottleneck. *Neural Computation*, 29(6):1611–1630, 2016. doi: 10.1162/NECO_a_00961. URL `https://doi.org/10.1162/NECO_a_00961`.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5, 2015. doi: 10.1109/ITW.2015.7133169. URL `https://doi.org/10.1109/ITW.2015.7133169`.

Naftali Tishby, Fernando Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999. URL `https://arxiv.org/abs/physics/0004057`.

James Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford, 2003.