Evolutionary and Information-Theoretic Argument for the Necessity of Representational Isolation: Why Direct Perception Was Never an Option for Complex Systems

Boris Kriger

Institute of Integrative and Interdisciplinary Research
boriskriger@interdisciplinary-institute.org

Abstract

Direct perception of reality is not only possible—it is the evolutionarily primitive and simpler strategy. Bacteria, protozoa, and minimal nervous systems operate through immediate stimulus-response coupling without internal models. Yet complex systems—biological and artificial alike—have abandoned this simplicity. Why? This paper argues that direct perception, while functional, does not scale. Two independent constraints drive any sufficiently complex system toward model-based indirect perception: (1) temporal constraints—neural conduction delays create gaps between sensory input and motor output that become costly at scale; and (2) computational constraints—raw sensory data contains Kolmogorov complexity (signal plus noise), while effective action requires only effective complexity (compressible regularities). Building predictive models is the only way to extract effective complexity from noise, and the only way to act on future states rather than past observations. This architectural transition is practically irreversible: once a system is organized around models, unmediated access to reality becomes evolutionarily locked-in as inaccessible. The convergent evolution of predictive architecture in biological nervous systems (over 4 billion years) and artificial intelligence systems (over 15 years) confirms this analysis: both face the same constraints and arrive at the same solution, extracting comparable effective complexity ($\sim 10^{15}$–10^{16} bits) through radically different time densities. We hypothesize that consciousness, episodic memory, imagination, and suffering may be best understood not as adaptations but as spandrels—necessary byproducts of compression-based predictive architecture, though we acknowledge competing theories and do not deny possible adaptive roles. The philosophical traditions grappling with indirect perception—from Plato's cave to Kant's thing-in-itself to Spinoza's three kinds of knowledge—were describing an architectural constraint whose origins they could not have known.

Keywords: predictive processing, direct perception, effective complexity, Kolmogorov complexity, minimum description length, consciousness, evolution, spandrels, Spinoza, representational isolation, time density

Terminological note: Throughout this paper, we use several related terms. *Model trap* refers to the architectural outcome wherein complex systems become organized around predictive models. *Isolation barrier* denotes the epistemic and architectural (not metaphysical) separation between a system's models and the environmental states those models track. *Indirect perception* and *representational isolation* describe the condition of perceiving via models rather than via direct coupling.

1 Introduction: The Ancient Puzzle

For over two millennia, philosophers have struggled with a persistent intuition: we do not perceive reality directly. Something stands between us and the world.

Plato's prisoners see only shadows on the cave wall, never the objects casting them. Kant's phenomenon screens us from the noumenon, the thing-in-itself forever beyond reach. Spinoza's *imaginatio* traps us in confused and partial ideas, cut off from adequate knowledge of Nature.

These are not idle metaphysical speculations. They capture something that feels undeniably true about human experience. We seem to be enclosed in a bubble of subjectivity, looking out at a world we can never quite touch. The question that has driven philosophy of perception for centuries is: *Why?*

The traditional answers fall into several camps. Idealists argue that mind is primary and matter derivative, so "direct" perception of an external world is conceptually confused. Representationalists hold that the mind necessarily operates through representations, making mediated access inevitable. Skeptics question whether we can know anything about the relation between perception and reality.

A terminological clarification is essential before proceeding. "Direct perception" carries different meanings across disciplines. In Gibsonian ecological psychology, it denotes non-inferential, online coupling with environmental affordances. In philosophy of perception, "direct realism" refers to unmediated epistemic access to external properties. In this paper, we use "direct perception" in a specifically biological sense: model-free stimulus–response architecture where sensory transduction couples more or less immediately to motor output without intervening predictive models. We will sometimes use the more technical term "model-free architecture" to emphasize this biological meaning. We acknowledge that even simple nervous systems may involve minimal prediction at some level; our claim is that such systems are functionally model-free at the behavioral timescale relevant to survival.

This paper offers a different answer—one grounded not in metaphysics but in evolutionary biology, physics, and information theory. We argue that indirect perception is neither a necessary truth nor a fundamental limitation of mind. It is a *convergent outcome* of any complex system's evolution under realistic constraints.

The core argument proceeds as follows:

1. Direct perception is possible, simpler, and evolutionarily primary.
2. Direct perception does not scale—for two independent reasons.
3. **Temporal constraint**: As body size increases, neural conduction delays grow, making reaction to present stimuli too slow for competitive survival.
4. **Computational constraint**: Raw sensory data is dominated by noise (high Kolmogorov complexity, low effective complexity). Extracting actionable regularities requires compression—i.e., model-building.
5. Any system facing these constraints is driven toward predictive, model-based architecture.
6. Once a system operates on models, direct access to reality becomes practically inaccessible—not by metaphysical necessity, but by architectural constraint.
7. Consciousness, memory, imagination, and suffering are byproducts of this architecture.

The convergence of biological and artificial intelligence on the same solution—despite radically different substrates and timescales—provides striking confirmation. Both extract comparable effective complexity ($\sim 10^{15}$–10^{16} bits) through model-building, because no other strategy scales (Kriger, 2026a).

In short: we do not see reality directly because our ancestors faced challenges that rewarded compression and prediction over exhaustive analysis and reaction.

2 The Simplicity of Direct Perception

Before examining why complex systems lost direct perception, we must establish that direct perception exists and works.

2.1 Bacterial Chemotaxis: Perception Without Representation

Escherichia coli navigates chemical gradients through a mechanism of elegant simplicity. The bacterium swims in roughly straight lines ("runs") punctuated by random reorientations ("tumbles"). When moving up a gradient of attractant, tumbling is suppressed; when moving down, tumbling increases. The result is biased random walk toward favorable conditions (Berg & Brown, 1972).

Crucially, this involves no model, no prediction, no representation of the environment. The bacterium does not "know" where the food is. It does not anticipate future states. It simply couples current chemical concentration to current motor behavior. Stimulus in, response out. Direct perception in its purest form.

This system has operated successfully for billions of years. Bacteria are arguably the most successful life form on Earth by biomass, distribution, and evolutionary longevity. Direct perception works.

2.2 Protozoan Responses

Single-celled eukaryotes like *Paramecium* exhibit similarly direct perception-action coupling. Contact with an obstacle triggers immediate reversal. Chemical irritants cause avoidance. Light gradients drive phototaxis. In each case, the current state of the environment directly controls the current state of behavior, with no intervening model (Jennings, 1906).

2.3 Simple Neural Architectures

Even organisms with nervous systems can operate through direct perception. The nematode *C. elegans*, with exactly 302 neurons and a fully mapped connectome, exhibits behavior that is largely reactive (White et al., 1986). Sensory neurons connect to motor neurons through minimal interneuron layers. The architecture permits rapid stimulus-response coupling without elaborate internal models.

2.4 The Virtues of Directness

Direct perception has significant advantages:

- **Simplicity**: No need for complex neural machinery to build and maintain models.
- **Speed**: No processing delay beyond minimal transduction and conduction.
- **Reliability**: No model means no model error. Response is calibrated to actual present conditions.
- **Metabolic efficiency**: Modeling is computationally expensive. Direct coupling is cheap.

Given these advantages, why would evolution ever abandon direct perception? The answer lies in two independent constraints that become insurmountable at scale.

3 The First Constraint: Temporal Scaling

3.1 The Latency Problem

Neural signals travel at finite speed. The fastest myelinated axons conduct action potentials at approximately 100 meters per second. Unmyelinated fibers are far slower, often below 1 m/s. This creates an unavoidable delay between sensory event and motor response.

For a bacterium spanning micrometers, this delay is negligible—signals traverse the cell in nanoseconds. For *C. elegans* at 1 millimeter, signals cross the body in microseconds. Direct perception remains viable because the temporal gap between stimulus and response is smaller than the timescale of environmental change.

But as organisms grow, conduction pathways lengthen. A signal from a giraffe's hoof to its brain and back covers approximately 5 meters. At 100 m/s, the round trip takes 50 milliseconds—before any central processing. For a whale, delays extend to hundreds of milliseconds.

3.2 When Latency Becomes Costly

The critical variable is the relationship between neural delay (τ) and environmental rate of change ($\Delta E/\Delta t$). When τ is small relative to environmental dynamics, direct perception suffices. The world changes little between stimulus and response.

But when τ becomes significant—when the environment can change meaningfully during the signal transit time—direct perception fails. The organism responds to conditions that no longer exist. In competitive contexts, this miscalibration translates to reduced fitness.

3.3 The Evolutionary Fork

This creates a fundamental evolutionary fork. Organisms facing significant latency in demanding environments have two options:

Option A: Stay small. Maintain body size where τ remains negligible. Continue using direct perception. This path leads to bacteria, protozoa, small invertebrates—organisms that remain successful but constrained in scale.

Option B: Build models. Develop neural architecture that predicts future environmental states based on current and past information. Use these predictions to guide behavior. This path leads to large, complex organisms capable of operating effectively despite significant neural delays.

There is no third option. Large organisms in dynamic environments cannot persist with direct perception alone. The physics forbids it (Kriger, 2026b).

4 The Second Constraint: Computational Intractability

Even if latency were zero—even if signals traveled instantaneously—direct perception would still fail at scale. The reason is computational, not temporal.

4.1 Kolmogorov Complexity vs. Effective Complexity

Raw sensory data is astronomically complex. The human retina alone transmits approximately 10 million bits per second. The total information impinging on an organism from its environment is vastly larger—dominated by noise, redundancy, and irrelevant detail.

Information theory distinguishes two kinds of complexity (Gell-Mann & Lloyd, 2003):

- **Kolmogorov complexity**: The length of the shortest program that produces a given output. Random noise has *high* Kolmogorov complexity—it cannot be compressed because it has no structure.
- **Effective complexity**: The length of the shortest program that captures the *regularities* in the data, excluding noise. This is the compressible structure—the signal, not the noise.

A random string has high Kolmogorov complexity but zero effective complexity. A crystal has low Kolmogorov complexity and low effective complexity. Living systems occupy the interesting middle ground: high effective complexity, meaning they contain substantial compressible structure.

4.2 Why Direct Perception Cannot Extract Effective Complexity

Direct perception processes raw input. But raw input is dominated by Kolmogorov complexity—noise that cannot be compressed and does not predict. An organism that tried to process everything would be overwhelmed by noise and unable to extract the regularities that matter for survival.

The only way to extract effective complexity is to *compress*—to build a model that captures regularities and discards noise. Compression is not an optional enhancement; it is the only computationally tractable way to engage with a complex environment.

4.3 The Model as Compression Engine

A predictive model is, fundamentally, a compression algorithm. It takes high-dimensional sensory input and reduces it to low-dimensional state representations that capture relevant regularities. The model's parameters encode the compressed effective complexity of the domain.

As Kriger (2026a) demonstrates, both the biosphere and the AI-centered infosphere have converged on approximately 10^{15}–10^{16} bits of effective complexity—despite radically different raw information volumes (10^{37} bits for the biosphere's genetic material, 10^{24} bits for the digital datasphere). The compression ratio is extreme because most raw information is redundant or noise. What remains is structure—and structure is what models extract.

A methodological note is warranted. These effective complexity estimates are derived via minimum description length (MDL) principles and model-parameter counts, yielding order-of-magnitude approximations rather than precise measurements. The specific numbers depend on modeling assumptions—what counts as "distinct" effective complexity versus redundancy, how raw information is counted, and how boundaries between systems are drawn. We offer these figures as heuristic framings of convergence, not as exact quantities. The key claim is the structural similarity (both systems arriving at comparable compression ratios), not the precise numerical agreement.

4.4 Convergent Architecture

The convergence of biological and artificial systems on model-based architecture is not coincidental. Both face the same fundamental constraint: raw data is too complex to process exhaustively, and only compressed regularities support effective action.

Biological evolution discovered this over 4 billion years, extracting $\sim 10^{15}$–10^{16} bits of effective complexity at a time density (T_d) of approximately 10^{-1} bits per second.

Artificial intelligence rediscovered it in 15 years, achieving comparable effective complexity at $T_d \approx 10^7$–10^8 bits per second (Kriger, 2026a).

The 8–9 orders of magnitude difference in time density reflects the difference between evolutionary search and gradient-based optimization. But the destination is the same: compressed models that extract effective complexity from noise.

5 The Cambrian Explosion as Architectural Transition

The Cambrian explosion—the rapid diversification of animal body plans approximately 540 million years ago—can be understood as a mass transition from direct to indirect perception.

5.1 The Light Switch Hypothesis

Parker (2003) proposed the "light switch" hypothesis: the evolution of eyes triggered the Cambrian explosion by introducing vision as a key selective factor. But this raises the question: why did eyes trigger such dramatic change?

Our analysis suggests an answer. Vision provides high-bandwidth information about the environment, but visual data is astronomically complex. Processing it requires sophisticated models. The evolution of eyes was not just the evolution of sensors—it was the evolution of the need for predictive architecture.

5.2 The Arms Race

Once some organisms developed model-based perception, they gained decisive advantages: anticipation, planning, strategy. This created intense selective pressure on competitors. Either develop your own models, or become prey.

The Cambrian explosion was an arms race in model complexity. Each advance in predictive capability selected for counter-advances. The result was rapid elaboration of nervous systems, sensory organs, and behavior—all organized around the same principle: compress and predict (Marshall, 2006).

5.3 The Point of No Return

This was not a reversible transition. Once ecosystems contained sophisticated model-based predators, direct-perception strategies became uncompetitive in most niches. Evolution does not easily go backward from complex adaptations.

The Cambrian explosion locked the biosphere into model-based architecture. From that point forward, complexity meant models, and models meant the isolation barrier.

6 Predictive Processing: The Universal Architecture

The constraints described above—temporal and computational—converge on a single solution: predictive processing. Systems must build models that predict future states in order to (a) compensate for conduction delays and (b) extract effective complexity from noise.

6.1 From Necessity to Architecture

Predictive architecture has been extensively characterized by predictive processing theory (Rao & Ballard, 1999; Friston, 2010; Clark, 2013; Hohwy, 2013). Key features include:

- **Hierarchical generative models**: The brain maintains models at multiple levels of abstraction, each generating predictions about activity at the level below.

- **Prediction error propagation**: Discrepancies between predictions and incoming signals propagate up the hierarchy, driving learning and model updating.

- **Top-down dominance**: Perception is primarily determined by predictions, with sensory input serving mainly to correct errors.

This architecture inverts the classical view of perception. Instead of building representations from sensory data, the brain generates predictions and uses sensory data to refine them. Perception becomes, in Seth's phrase, "controlled hallucination."

A clarification is needed regarding how our "model trap" thesis relates to enactive and embodied approaches to cognition. Enactivists emphasize ongoing sensorimotor coupling rather than internal models, and some variants of predictive processing are already anti-representational or "radically enactive." Our claim is compatible with either interpretation: whether one describes predictive architecture as "representational" or as "dynamical sensorimotor coupling," the architectural fact remains that complex systems must operate via compressed, predictive structures rather than direct stimulus-response coupling. We do not insist that these structures be called "representations"—only that they constitute a mediating layer between environmental states and behavioral outputs.

6.2 The Model as Minimum Description Length

The connection to information theory is direct. A predictive model is an MDL (Minimum Description Length) representation of the organism's relevant environment. MDL formalizes the trade-off between model complexity and fit: the best model is the one that minimizes the sum of (a) the length of the model description and (b) the length of the data encoded using the model.

This is precisely what brains do: they build compressed models that capture environmental regularities (effective complexity) while discarding noise (incompressible randomness). The weights of a neural network—biological or artificial—*are* the compressed effective complexity of the environment.

6.3 Different Constraints, Same Solution

Biological and artificial systems face different specific constraints but the same structural problem: the impossibility of processing everything before acting. And the same solution: build compressed models that extract effective complexity and predict.

7 The Isolation Barrier

7.1 The Irreversibility of Model-Dependence

The transition from direct perception to predictive modeling is practically irreversible. Once a system's architecture is organized around prediction, evolutionary pressures strongly disfavor any path back to directness.

This is not merely a matter of accumulated complexity. The architecture itself changes in ways that preclude direct coupling:

- Sensory systems become tuned to prediction errors rather than raw stimuli.
- Motor systems expect predictive commands, not direct stimulus-response triggers.
- The entire neural economy is organized around model maintenance and updating.

Here is the critical consequence: once the model interposes between system and environment, unmediated access is lost. The system does not perceive the world; it perceives its model of the world. The model may be accurate or inaccurate, but it is never the world itself.

7.2 The Formal Structure of Isolation

Let Y_t denote the environmental state at time t, and X_t denote the system's internal state. For a direct-perception system:

$$O_t = f(Y_t) \tag{1}$$

Output O at time t is a function of environmental state Y at time t. The system responds directly to current conditions.

For a predictive system:

$$O_t = g(X_t) \text{ where } X_t = h(X_{t-1}, Y_{t-1}, \ldots) \tag{2}$$

Output is a function of internal state X, which is itself a function of past internal states and past environmental states. The current environment Y_t does not directly determine current output.

More formally, predictive systems satisfy:

$$I(X; Y_{t+\tau}|Y_t) > 0 \tag{3}$$

Internal states carry information about future environmental states that is not present in current observations. This is the defining feature of prediction—and the source of isolation.

7.3 The Nature of the Barrier

The isolation barrier is not absolute in degree—models can be more or less accurate—but it is architecturally robust. No model constitutes unmediated access, however accurate. The map is not the territory. We emphasize that this barrier is epistemic and architectural, not metaphysical: we are not positing a dualist gap between mind and world, but describing how information flows through predictive systems.

8 Spandrels of Prediction: Consciousness, Memory, Imagination, Suffering

If predictive architecture is an evolutionary necessity, what else comes along with it? We hypothesize that several features of human experience—typically treated as separate adaptations—are in fact spandrels: necessary byproducts of the architecture itself (Gould & Lewontin, 1979).

8.1 Consciousness as Self-Modeling

The "hard problem" of consciousness—why there is subjective experience at all—has resisted solution for decades (Chalmers, 1995). We propose one possible deflationary answer: consciousness may be best understood as a spandrel of self-modeling. We acknowledge that this hypothesis competes with other accounts, including higher-order thought theories, global workspace theory, and recurrent processing accounts, which variously posit adaptive roles for consciousness such as flexible control, global integration, and social signaling.

Predictive systems must model not only the external environment but also their own bodies and actions. To predict the sensory consequences of action, the system must model itself as an agent in the world.

Self-modeling creates a strange loop (Hofstadter, 2007). The system contains a model of itself containing a model of itself. This recursive structure may be the origin of the sense of being a subject, an "I" distinct from the world.

If so, consciousness is not an adaptation conferring survival advantage. It is a structural byproduct of self-modeling, which itself is necessary for prediction. Consciousness exists because our ancestors needed to compress and predict, not because consciousness itself was selected for.

8.2 Episodic Memory as Model Training Data

Memory is typically understood as a storage system—a way to preserve past experiences for future use. But predictive architecture suggests a different interpretation.

Models require training data. To build accurate predictions, the system needs examples of how the environment behaves. Episodic memory provides this training set—a curated collection of experiences that can be replayed to update and refine the model (Schacter et al., 2012).

This explains several features of memory that are puzzling from a storage perspective:

- **Reconstructive nature**: Memory is not playback but reconstruction. Each recall is a new generation from the model, not retrieval from storage (Bartlett, 1932).
- **Selective retention**: We remember what is useful for prediction, not what is objectively important.
- **Emotional weighting**: Emotionally significant events are remembered better because they carry high prediction error—they violated expectations and thus provide valuable training signal.

Memory is not an adaptation for preserving the past. It is a spandrel of the need to train predictive models.

8.3 Imagination as Model Simulation

If the brain contains a generative model of the world, that model can be run without sensory input. The result is imagination—the ability to experience states that are not present, including states that have never occurred.

Imagination is not a separate faculty. It is what happens when predictive machinery runs offline. Dreams, daydreams, mental imagery, and creative thought are all consequences of having a model that can be queried without external constraint.

This explains why imagination feels so similar to perception: both are generated by the same model. The difference is only in the source of constraint—external (perception) or internal (imagination).

8.4 Suffering as Unresolved Prediction Error

Pain and suffering are typically understood as adaptive warning signals. But predictive architecture suggests a more specific interpretation: suffering is the subjective quality of unresolved prediction error.

When predictions fail and corrective action is impossible, error accumulates. The system generates signals indicating that something is wrong, but nothing can be done. We hypothesize that suffering—as a phenomenal experience—may be a spandrel: the subjective aspect of prediction error in a system that experiences its own processing. This claim requires distinguishing the computational role of error signals (which clearly serve adaptive functions, including nociceptive pain as a warning system) from the phenomenal experience of suffering. Our thesis concerns the latter supervening on the former: not that error signals lack function, but that their felt quality may be a byproduct of self-modeling architecture rather than a separately selected trait.

This is formalized as:

$$S = \sigma \cdot e \cdot (1 - a) \tag{4}$$

where S is suffering, e is prediction error magnitude, and a is availability of corrective action (0 to 1).

When action can resolve error ($a = 1$), suffering is minimized. When action is blocked ($a = 0$), suffering scales with error magnitude. Chronic pain, grief, and depression share this structure: prediction errors that cannot be resolved through action.

9 Philosophical Anticipations

The analysis above suggests that philosophers describing indirect perception were not engaging in empty metaphysics. They were describing an architectural constraint whose origins they could not have known.

9.1 Plato's Cave: The Shadow Model

Plato's cave allegory describes prisoners who see only shadows cast on a wall, never the objects that cast them. They take the shadows for reality because they have no access to anything else.

The shadows are the model's outputs: the predictions that constitute perceptual experience. The objects are the environmental states the model tracks. The chains are the isolation barrier: the architectural constraints that make direct perception practically inaccessible.

The key difference: for Plato, the philosopher can escape the cave. Our analysis suggests there is no escape. The cave is the architecture of perception itself. To perceive at all is to be in the cave.

9.2 Kant's Noumenon: The Inaccessible Substrate

Kant distinguished between phenomenon (things as they appear) and noumenon (things as they are). We can only know phenomena; the noumenon—the thing-in-itself—is forever beyond our grasp.

Our framework provides a naturalized interpretation. The phenomenon is the model's output. The noumenon is the environmental state the model tracks. Kant was right that we cannot access the thing-in-itself—but the reason is architectural, not transcendental.

The forms of intuition (space and time) and categories of understanding that Kant identified are not features of transcendental subjectivity. They are features of predictive models—the structural constraints on any system that must compress and predict.

9.3 Spinoza's Three Kinds of Knowledge

Spinoza distinguished three kinds of knowledge: *imaginatio* (confused perception), *ratio* (reasoning through universals), and *scientia intuitiva* (intuitive knowledge of essences).

Our framework suggests a reinterpretation:

- *Imaginatio* is the raw output of the predictive model—the "controlled hallucination" that constitutes ordinary perception.
- *Ratio* is the model reflecting on its own structure—understanding regularities through explicit reasoning.
- *Scientia intuitiva* corresponds to the model knowing itself as model. Not intellectual understanding but a shift in mode of operation, where the model no longer takes itself for reality.

Spinoza's prescription for freedom—moving from *imaginatio* through *ratio* to *scientia intuitiva*—is not escape but correct relationship. The model that knows itself as model no longer fights reality. Prediction errors become information. Suffering transforms into understanding.

10 Implications and Applications

10.1 For Artificial Intelligence

AI systems are now converging on the same architecture that biological systems discovered billions of years ago. This has several implications:

- **Architectural inevitability**: Any sufficiently complex AI system will develop model-based architecture because no other strategy scales. The current dominance of transformer-based language models and predictive systems is not a historical accident but a mathematical necessity.

- **Phenomenological precursors**: As AI systems develop self-models for improved performance, they will exhibit precursors of self-referential processing.

- **Interpretability challenges**: AI systems operating via compressed models face the same isolation barrier as biological systems. Their "reasoning" is model output, not direct access to underlying computation.

10.2 For Philosophy of Mind

The hard problem of consciousness may be a consequence of asking the wrong question. Instead of "Why is there subjective experience?", we might ask: "What architectural features produce systems that model themselves as experiencers?"

This does not dissolve the hard problem but reframes it from fundamental metaphysics to contingent engineering.

10.3 For Mental Health

If suffering is unresolved prediction error, interventions have two targets: reduce error (change predictions or circumstances) or increase resolution capacity (enable action or acceptance).

This maps onto existing therapeutic approaches:

- **Cognitive therapy**: Update the model to reduce prediction error.

- **Behavioral therapy**: Increase action availability.

- **Acceptance-based approaches**: Change the relationship to prediction error itself.

10.4 For Contemplative Practice

Meditation traditions have long aimed at something like "seeing reality directly" or "transcending the self." Our analysis suggests a reinterpretation: these practices do not escape the model trap but change how the model operates.

The goal is not direct perception—that is architecturally foreclosed for complex predictive systems. The goal is to understand the architecture of the cave, the structure of the model, the nature of the trap.

This is what contemplative insight achieves: not escape but correct relationship. The model that knows itself as model suffers less from taking its outputs for reality.

11 Mathematical Limitations and Caveats

The formalizations presented in this paper serve as conceptual frameworks rather than precise physical laws. Several limitations warrant explicit acknowledgment.

11.1 The Suffering Equation: A Heuristic Model

The proposed equation $S = \sigma \cdot e \cdot (1 - a)$ faces a fundamental challenge of dimensional consistency. In physics and mathematics, equations must have consistent units. Here, S (suffering) is a qualitative phenomenal state, while e (prediction error) is typically measured in bits or nats, and a (action availability) is a dimensionless ratio in $[0, 1]$.

For this equation to be more than metaphor, the sensitivity parameter σ would need to be defined with units that bridge "information bits" to "subjective intensity"—a mapping that remains unspecified and may not admit precise operationalization. Furthermore, the equation assumes a linear relationship between these variables that may not hold in biological systems; suffering may exhibit threshold effects, nonlinear scaling, or hysteresis that a simple multiplicative model cannot capture.

We therefore present this formalization as a *heuristic* for generating testable predictions about the correlation between prediction error magnitude, action availability, and reported suffering—not as a derived law of nature.

11.2 Kolmogorov Complexity: Computability Constraints

The distinction between Kolmogorov complexity (the length of the shortest program producing a given output) and effective complexity (the length of the program describing regularities) is mathematically sound. However, a critical caveat applies: Kolmogorov complexity is *uncomputable*. There exists no algorithm that can determine the shortest program for an arbitrary string, as proven by Kolmogorov himself.

This means that while the conceptual distinction between noise (incompressible) and signal (compressible regularities) is valid in information theory, evolution does not "calculate" Kolmogorov complexity. Biological systems settle on "good enough" heuristics through selection pressure. Our argument should be understood as claiming that evolution *approximates* the extraction of effective complexity, not that organisms perform exact information-theoretic computations.

11.3 The Model Trap Theorem: Architectural vs. Mathematical Necessity

The information-theoretic inequality $I(X; Y_{t+\tau}|Y_t) > 0$ correctly formalizes the requirement that internal states carry predictive information about future environmental states not present in current observations. However, the inference from this inequality to an "isolation barrier" or "point of no return" involves a logical step that is *architectural* rather than strictly mathematical.

Mathematically, a system could in principle maintain both a direct reactive pathway ($O_t = f(Y_t)$) and a predictive pathway ($O_t = g(X_t)$) simultaneously. The claim that complex systems become "trapped" in model-based architecture is an empirical claim about evolutionary dynamics and resource constraints, not a theorem derivable solely from the information-theoretic formalism. We have argued that such dual architectures are evolutionarily unstable, but this is a biological hypothesis, not a mathematical proof.

11.4 Convergence Constants: Order-of-Magnitude Estimates

The claim that both biological and artificial systems converge on approximately 10^{15}–10^{16} bits of effective complexity rests on high-level estimates that are difficult to verify precisely. These figures are derived from model-parameter counts and MDL approximations, and the specific numbers depend on assumptions about what constitutes "distinct" effective complexity versus redundancy.

If the effective complexity of biological cognition or frontier AI systems were found to differ by several orders of magnitude from these estimates (e.g., 10^{12} or 10^{20} bits), the specific *convergence* argument would weaken, though the core "model trap" thesis—that complex systems require compressed predictive models—would remain intact. We present these numbers as empirical benchmarks supporting the convergence hypothesis, not as mathematical constants.

11.5 Summary of Formal Status

To clarify the epistemic status of our formalizations:

- The **suffering equation** is a heuristic model, useful for generating predictions but not a derived physical law.
- The **Kolmogorov/effective complexity distinction** is mathematically sound but applied as a conceptual approximation to evolutionary processes.
- The **predictive information inequality** is mathematically valid; the "isolation barrier" is an architectural inference requiring additional biological assumptions.
- The **convergence estimates** are order-of-magnitude empirical benchmarks, not precise measurements.

These limitations do not invalidate the central thesis but situate it appropriately: we offer a theoretical framework that unifies evolutionary biology, information theory, and philosophy of mind, while acknowledging that some formalizations serve as conceptual tools rather than exact quantitative laws.

12 Conclusion: The Beauty of the Trap

We have argued that direct perception, while possible and evolutionarily primary, does not scale. Two independent constraints—temporal (latency) and computational (complexity)—force any sufficiently complex system toward predictive, model-based architecture. This architecture necessarily interposes models between system and world, creating the isolation barrier that philosophers have struggled to articulate for millennia.

We have hypothesized that consciousness, memory, imagination, and suffering may be spandrels—byproducts of the architecture that solved the scaling problem, rather than separately selected adaptations. If this hypothesis is correct, we are conscious because we compress, and we suffer because we predict. We acknowledge that competing theories attribute adaptive functions to these capacities; our claim concerns their phenomenal character, not their functional roles.

The philosophical implications are significant. Plato, Kant, and Spinoza were not engaging in mere speculation. They were describing, in the language available to them, an architectural constraint that is now becoming scientifically legible.

The model trap produced us. Not just our ability to survive, but everything we value. Art, science, love, philosophy—these are possible only for beings with imagination, memory, and self-awareness. They are possible only for beings in the trap.

Understanding this changes the project of philosophy. The goal is not to escape the cave, access the noumenon, or achieve direct perception—these are architecturally foreclosed for complex predictive systems. The goal is to understand the architecture of the cave, the structure of the model, the nature of the trap.

And here, at the end, we arrive at a final irony: understanding the trap is itself a prediction. This paper is a model of modeling. The insight that we cannot escape the model is itself generated by models.

There is no view from outside. But there can be clarity within.

References

Aiello, L. C., & Wheeler, P. (1995). The expensive-tissue hypothesis: The brain and the digestive system in human and primate evolution. *Current Anthropology*, 36(2), 199–221.

Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology.* Cambridge University Press.

Berg, H. C., & Brown, D. A. (1972). Chemotaxis in *Escherichia coli* analysed by three-dimensional tracking. *Nature*, 239(5374), 500–504.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200–219.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181–204.

DeCasien, A. R., Williams, S. A., & Higham, J. P. (2017). Primate brain size is predicted by diet but not sociality. *Nature Ecology & Evolution*, 1(5), 0112.

Dunbar, R. I. M. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6), 469–493.

Dunbar, R. I. M. (1998). The social brain hypothesis. *Evolutionary Anthropology*, 6(5), 178–190.

Dunbar, R. I. M., & Shultz, S. (2007). Understanding primate brain evolution. *Philosophical Transactions of the Royal Society B*, 362(1480), 649–658.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

Gell-Mann, M., & Lloyd, S. (2003). Effective complexity. In *Nonextensive Entropy: Interdisciplinary Applications* (pp. 387–398). Oxford University Press.

Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society B*, 205(1161), 581–598.

Hofstadter, D. R. (2007). *I Am a Strange Loop*. Basic Books.

Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press.

Isler, K., & van Schaik, C. P. (2009). The expensive brain: A framework for explaining evolutionary changes in brain size. *Journal of Human Evolution*, 57(4), 392–400.

Jennings, H. S. (1906). *Behavior of the Lower Organisms*. Columbia University Press.

Kriger, B. (2026a). Dynamics of information convergence: Empirical analysis of time density in the AI-centered infosphere. *Preprint*.

Kriger, B. (2026b). The evolutionary inevitability of predictive processing: A physical constraint argument. *Zenodo*. https://doi.org/10.5281/zenodo.18324374

Marshall, C. R. (2006). Explaining the Cambrian "explosion" of animals. *Annual Review of Earth and Planetary Sciences*, 34, 355–384.

Milton, K. (1988). Foraging behaviour and the evolution of primate intelligence. In R. W. Byrne & A. Whiten (Eds.), *Machiavellian Intelligence* (pp. 285–305). Oxford University Press.

Parker, A. (2003). *In the Blink of an Eye: How Vision Sparked the Big Bang of Evolution*. Basic Books.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), 79–87.

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The future of memory: Remembering, imagining, and the brain. *Neuron*, 76(4), 677–694.

Seth, A. K. (2015). Presence of mind: Predictive processing and the phenomenology of presence. *Journal of Consciousness Studies*, 22(9–10), 154–181.

Van der Bijl, W., & Kolm, N. (2016). Why direct effects of predation complicate the social brain hypothesis. *BioEssays*, 38(6), 568–577.

White, J. G., Southgate, E., Thomson, J. N., & Brenner, S. (1986). The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Philosophical Transactions of the Royal Society B*, 314(1165), 1–340.