The Imagination Machine: A Fixed Point in Human Knowledge

Orientation to a Framework for Embedded Epistemic Systems

Mark Tracy

Close your eyes.
Imagine your body positioned exactly how it is — only it's floating in front of you.
Now imagine a bubble around that body.
Now realize that you have become the surrounding darkness—
the outer boundary around that bubble.
You are a vanishing point of perspective:
a view from somewhere that appears nowhere.

1 Introduction

This document orients the reader to the Imagination Machine series, which consists of ten papers plus this introductory orientation, developing a unified formal framework for embedded epistemic systems. A reasoning system is one that stabilizes and utilizes representations of relations. An epistemic system is a reasoning system whose representational processes themselves become objects of analysis, supporting stabilization of internal relations and recursive modification thereof. An embedded epistemic system is completely contained within the world it is attempting to represent.

The series is written retrospectively — from the vantage point of a completed arc rather than a projected one. It was not designed in advance. It arrived all at once, in a moment's intuition that radically and utterly destabilized my experience; and it stabilized into its present form through the same recursive process it describes. This preface is therefore not a map drawn before the journey but a description of the territory as it emerged.

An embedded epistemic system can at most classify the ways in which it classifies the world, within the world itself. We begin, then, from an ontological ground that is neither external to the embedded epistemic system nor an internal capability of the system, but rather the precondition for internality and externality to arise at all: Paper 0 argues that demarcation and abstraction — the co-arising operations of differentiation-without-division and unification-without-annihilation — are ontologically prior to time and space, and names their shared primitive unity-in-difference. From that ground the series moves through formal epistemology, the philosophy of science, a treatment of systems and prediction, the structure of analogy and abstraction, and a personal theological note — arriving, in Paper VIII, at the phenomenological grounding that explores a view from somewhere inside the architecture, before closing, in Paper IX, on the moral consequence that the formal framework cannot derive but can only demand.

2 The Core Epistemic Loop

The central operation of the framework can be summarized as the following cycle.

1. An agent observes data generated by interaction with an environment.
2. Observations are compressed into a representation — a world model — that retains relational invariants while discarding detail.
3. The compressed representation is extended through the implication of missing relations.
4. Upon taking an implied relation as a basis for action, inconsistency between expectation and observation drives revision of the world model.

Repeated execution of this loop gradually stabilizes world models that capture persistent relational structure in the environment. Such stabilized structures function operationally as knowledge. Self-consistent world models appear as fixed points of the operator induced by this loop: models whose own implied observational profiles, when reinterpreted through inference, reproduce the models themselves.

This perspective resonates with several research traditions in which learning is understood as a dynamical feedback process. Early cybernetic work emphasized the centrality of feedback loops in adaptive systems [11, 1]. More recent work in neuroscience proposes predictive processing models in which perception and cognition arise through the minimization of prediction error [4, 2]. Reinforcement learning frameworks likewise describe agents that iteratively update internal models based on interaction with their environment [10].

The framework also bears philosophical affinity with Karl Popper's conception of knowledge growth through conjecture and refutation [7, 8, 9]. Within the present framework, extension operations generate candidate structural hypotheses, while prediction error functions as a mechanism of selective elimination guiding representational revision.

3 Representation and Closure

A central philosophical challenge for embedded epistemic systems is that representation necessarily involves the imposition of conceptual boundaries upon a world that cannot be accessed independently of those boundaries.

Hilary Lawson has argued that all representation involves acts of closure through which distinctions are drawn and stabilized [6]. The present framework formalizes this picture: the inference–implication loop is the closure mechanism; the fixed points of the operator it induces are the stable closures; the quotient space Q_w is the structured texture through which an embedded system encounters the world under model w. Compression and representation are not two operations but one: to compress observations into a quotient is to represent them, and to represent them is to have imposed a closure.

A key structural feature is that classifiers themselves belong to the observation space. This follows from the conditions of self-representation: any system capable of epistemic reasoning must be able to encounter and revise its own acts of classification. As a result, the evaluative processes that guide model selection — valuation and will — also appear as observable elements subject to the same representational compression. Will is not explained away by the framework; it is what remains when the inference–implication loop has done everything it can do.

4 A Layered Architecture

Although the papers in the series address diverse domains, they can be viewed as exploring different layers of a single architecture.

- **Ontological Ground.** Paper 0 proposes that the physical notions of time and space are ontologically posterior to the co-arising notions of demarcation and abstraction. To demarcate is to hold difference atop unity; to abstract is to hold unity atop difference. Neither is prior to the other. Their shared primitive — unity-in-difference — is what must be in place for questions of ordering, locating, or relating to arise at all. The formal operations of the series presuppose this primitive at every level.

- **Epistemic Foundations.** Paper I examines the situation of an embedded observer and introduces the inference–implication loop through which world models stabilize as fixed points. The world model is the quotient graph induced by compression of the observation space; physical laws appear as relational invariants in this quotient; entropy arises as a measure of the compression itself. The paper explicates that classifiers belong to the observation space — and locates will as the irreducible remainder of the loop.

- **Philosophy of Science.** Paper II interprets scientific knowledge as the stabilization of relational invariants under compression of observational data, and identifies reproducibility as the condition that two observers' quotient structures agree on the preserved invariants.

- **Dynamical Systems.** Paper III develops a general formalism for systems and agent–environment coupling, providing the dynamical structure within which the representational closures of Paper I arise for embedded epistemic systems.

- **Prediction and Representational Closure.** Paper IV develops the quasi-periodic environment as the naturalistic setting in which human temporal metacognition evolved, and formalizes a minimal predictive agent that recovers latent dynamical structure through prediction error alone. The Koopman eigenfunction structure of the relational observables becomes linearly recoverable through prediction error alone, under a fixed architectural constraint. The paper generalizes this result to arbitrary sequences over finite dictionaries via a canonical cyclic lifting: any sequence over an alphabet of size M is indexed by $\mathbb{Z}/M\mathbb{Z}$, and pairwise modular differences lifted to the unit circle produce an observation vector the architecture processes without modification. Interestingly, the linear encoding of the Koopman spectrum is found to be distributed throughout the full converged parameter vector: randomly drawn same-size slices of the parameter vector recover the Koopman spectrum with near-identical fidelity across independent draws, establishing that next-step prediction training organizes the full parameter vector as a linear function of the Koopman spectrum rather than concentrating it at any privileged location.

- **Analogy and Abstraction.** Paper V introduces a formal account of analogy, arguing that any analogy is mediated by an abstract domain of which both source and target are instances. Building on Gentner's structure-mapping theory [5], the paper shows how abstraction enables structural transfer across domains.

- **Horn Filling and Self-Demonstration.** Paper VI identifies a common structural pattern — the extension schema — across holonic composition, simplicial horn filling, and analogical abstraction, and demonstrates that the act of identifying this correspondence is itself a fourth

instantiation of the schema. The argument exhibits the structure it analyzes: the reader watches the schema execute.

- **Geometric Theology.** Paper VII is a personal note on the intuition underlying the series. It identifies the three-sphere — whose center is inaccessible and everywhere equidistant from within its surface volume — as an epistemically honest and coherent cosmology, and it draws historical parallels between such a view and ancient theological traditions.
- **Phenomenological Grounding.** Paper VIII examines the semiotic constitution of the embedded observer: perception is always already meaning-laden, no organism has access to a view from nowhere, and the apparent triad of faith, logic, and experience resolves into a single recursive loop — the phenomenological form of the inference–implication loop.
- **Moral Philosophy.** Paper IX extends the framework to the domain of moral action. Will appears as the irreducible remainder of the inference–implication loop; the paper formalizes what it means for that choice to be morally admissible, proposing an augmentation of Kant's Categorical Imperative in which the object of universalization is not an action alone but a tuple of action and motivation set.

5 Reading the Series

The papers of the Imagination Machine series may be read independently, but they collectively describe different aspects of the same architecture.

Readers interested primarily in the formal epistemological framework may begin with Paper I, which defines the inference–implication loop, argues the inclusion $C \subseteq D$, and locates will as the loop's irreducible remainder. Paper II applies this framework to the philosophy of science.

Readers interested in the dynamical and computational dimensions of the framework may focus on Papers III and IV, which develop the agent–environment formalism and the predictive agent in quasi-periodic environments, together with the generalization to arbitrary symbolic sequences via cyclic lifting.

Readers interested in the structure of reasoning itself may go directly to Papers VI and VII, which develop the theory of analogy, abstraction, and the extension schema. Paper VI's self-demonstration is the series' most explicit statement of its own recursive structure.

Paper VII may be read at any point after the central architecture of the series has become visible. It requires no technical background. It is the series looking at itself from the only vantage point its author has: from inside.

Paper VIII provides the phenomenological grounding for the framework and may be read first or last. Read first, it prepares the reader to inhabit the formal structure rather than merely observe it. Read last, it shows what the formal structure felt like from the inside throughout.

Paper IX closes the series on the moral demand the framework cannot itself supply. It may be read at any point, and it ought to be read.

6 Conclusion

The series ultimately argues that the constraint of an embedded epistemic system that can at most classify the ways in which it classifies the world, within the world itself, is not merely a limitation of knowledge but the condition under which knowledge becomes possible at all. A system that

cannot step outside its own representational frame does not thereby fail to know; it knows in the only way knowing is possible: from somewhere.

The view from nowhere is not available. This is not a deficiency. It is what makes the view from somewhere invaluable.

The moral consequence of the same structure is not a derivation but a demand. At the point where the inference–implication loop exhausts itself, will remains — and will cannot be derived. What the framework stipulates is only that this remainder is irreducible. What is done with it must be chosen. The foremost moral principle is therefore to seek moral principles. Paper IX formalizes what it means for that seeking to be admissible — but the seeking itself must come first, and it comes from nowhere the loop can reach.

The bubble, in the end, was never just a metaphor: it was the containing structure. The view from somewhere that appears nowhere, contained within its unknowable whole, is not a failure of perspective: it is the necessary condition of any perspective at all.

References

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

[2] Andy Clark. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press, 2016.

[3] Brendan Fong and David I. Spivak. *Seven Sketches in Compositionality: An Invitation to Applied Category Theory*. Cambridge University Press, 2019.

[4] Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

[5] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983.

[6] Hilary Lawson. *Closure: A Story of Everything*. Routledge, 2001.

[7] Karl R. Popper. *The Logic of Scientific Discovery*. Hutchinson, 1959.

[8] Karl R. Popper. *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge, 1963.

[9] Karl R. Popper. *Objective Knowledge: An Evolutionary Approach*. Oxford University Press, 1972.

[10] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

[11] Norbert Wiener. *Cybernetics: Or Control and Communication in the Animal and the Machine*. MIT Press, 1948.

The Imagination Machine, Paper 0: What is Prior to Time and Space?

Mark Tracy
Boston University
mrktracy@bu.edu

May 2026

The physical notions of time and space—whether understood phenomenologically or theoretically—are ontologically posterior to the co-arising notions of demarcation and abstraction.

By saying that one thing A is ontologically prior to another thing B, I mean that A is necessary for B to be intelligibly conceived at all. It is equivalent to say that B is ontologically posterior to A. For example, light is ontologically prior to a shadow.

Demarcation and abstraction are ontologically prior to time and space. To have demarcated something is to have differentiated—without dividing—what is otherwise a unity; demarcation is thus the holding of difference atop unity. Abstraction, on the other hand, is the association of differentiated instances with a common representation; abstraction, then, is the holding of unity atop difference.

One conception of time and space is as orthogonal axes of reality that index events. This conception presupposes the notion of extent (for example, Einsteinian spacetime presupposes a metric structure), which in turn presupposes demarcation insofar as something has extent if it may be demarcated. Carrying on the illustrative example, a metric structure presupposes a topology, with its attendant notions of closure and openness that formalize a notion of demarcation, since the very possibility of "open" subsets with "closed" complements presupposes the intelligibility of complementary distinction within a whole—that is, demarcation of what is held at once to be a unity.

Inherent in the sensibility of demarcation is the sensibility of abstraction. That is, to hold unity at once as differentiated is to hold difference at once as unified. Demarcation and abstraction are therefore co-dependent concepts, each relying on the other for its own intelligibility. They are ontologically co-arising—neither being prior to the other—and may be understood as different orientations of the same primitive. If we dare give it a name, let us call this primitive "unity-in-difference."

The co-dependence of demarcation and abstraction is illustrated by our prior example: inherent in the definition of a topology is the abstractive capacity to associate "elements" into a common higher-order representation called a set; this, in turn, relies upon a notion of demarcation for the sensibility of differentiated "elements" at all.

Having traced this chain of dependency to its co-dependent generative concepts, we conclude that demarcation and abstraction ontologically precede both time and space. For example, the duality of "before" and "after" is not temporally primitive but demarcationally primitive: only with a commitment to difference held atop unity, and unity held atop difference, does a relational ordering of states become intelligible at all. In other words, "time" is ontologically posterior to demarcation in something that is nonetheless held to be one and the same object.

Similarly, the duality of "here" and "there" is not spatially primitive but abstractly primitive: only with a commitment to unity held atop difference, and difference held atop unity, can such relational objects as "here" and "there" be intelligibly conceived. In this view, "space" is ontologically posterior to demarcation and abstraction in the following sense: any distance is necessarily between differentiated relata, relative to a reference frame—that is, an observer, a third differentiated relatum. It is precisely the unity that contains all such relations and relata that we call "space."

Taken together, these considerations suggest that time and space are not ontological primitives but rather rely for their intelligibility upon a more basic notion of unity and difference held together without collapse to either pole. Demarcation and abstraction—understood as co-arising orientations of this single primitive that we have called unity-in-difference—are necessary for the intelligibility of temporal ordering and spatial relation, just as light is necessary for the intelligibility of shadow. This is not to say that time and space are illusory or dispensable; they remain indispensable constructs within their proper domains. But they are posterior in the sense that they presuppose a prior structure of differentiation-without-division and unification-without-annihilation. To ask what is prior to time and space is thus not to ask what "came before" them, or what is "beyond" or "behind" them, but to ask what must exist in order for questions of time, space, ordering, locating, or relating to arise at all.

The Imagination Machine I: A View from Somewhere

Epistemic Closure, Physical Law, and Entropy Embedded in a Block Universe

Mark Tracy
Boston University
mrktracy@bu.edu

Abstract

This paper develops a minimal formal framework for epistemology under the constraint that epistemic systems are embedded within the world they attempt to model. Because such systems lack access to an external vantage point, knowledge cannot be defined by correspondence with an independently accessible reality. Instead, epistemic coherence must arise from internal structural consistency.

Observations generate world models through an inference map, while world models generate canonical observational profiles through an implication map. Together these maps form an inference–implication loop that induces an operator on model space. Self-consistent world models appear as fixed points of this operator: models whose own implied observational profiles, when reinterpreted through inference, reproduce the models themselves. Each model therefore acts as a compression of the observation space, inducing a classifier and a corresponding quotient representation of observations.

A key structural feature of the framework is that classifiers themselves belong to the observation space. This follows from the conditions of self-representation: any system capable of epistemic reasoning must be able to encounter and revise its own acts of classification. As a result, the evaluative processes that guide model selection—valuation and will—also appear as observable elements subject to the same representational compression.

Within a given model, empirical regularities emerge as relational invariants in the induced quotient space, while entropy arises as a measure-theoretic quantity associated with the same compressive structure. The framework therefore characterizes scientific theories as stable representational compressions of observational structure for agents embedded within the environments they model.

1 Introduction

Embedded epistemic systems cannot access the universe from outside. Observations, models, classifiers, and their relations therefore exist as structures within the same universe. No external vantage point is available from which to define correspondence between representation and world.

The guiding constraint is:

> An embedded epistemic system can at most classify the ways in which it classifies the world, within the world itself.

Rather than describing temporal learning, we treat the universe as a single relational structure containing observations, models, and consistency relations between them. Within such a framework

coherence must be defined internally, as the closure of the inference–implication loop rather than as external correspondence.

This position is closest in spirit to Hilary Lawson's closure theory of the world. Lawson argues that openness—raw, unstructured reality—is fixed as "something" only through interventions he calls closures, and that no closure fully captures the openness beneath it. The present framework formalizes a version of this picture. The inference–implication loop is the closure mechanism; the fixed points of the operator it induces are the stable closures; a quotient space Q_w is the closed texture through which an embedded system encounters the world under the model w. The crucial point is that what a model implies is not best understood as a single isolated observational consequence, but as a canonical observational profile internal to that closure: a structured way the world shows up for a life situated within the model.

But the framework adds something to Lawson's account that his descriptive language leaves implicit: the acts of will and valuation that select among possible closures are not external to the representational structure. Because classifiers are themselves observations—for reasons derived in Section 3 rather than merely asserted—valuation is interior to the system it animates. This is the structural heart of the paper.

Two clarifications are important at the outset. First, the framework is not a form of coherentism in which any internally consistent system of representations counts as knowledge. The structure of observations within the universe constrains admissible models through the probability measure introduced below. Closure of the inference–implication loop occurs only relative to this observational structure. Second, the framework does not deny the existence of an external world. It instead observes that embedded epistemic systems cannot compare representations with that world directly. The problem addressed here is therefore structural rather than metaphysical.

A further clarification concerns model-relativity. Different self-consistent models may in principle induce different quotient spaces and therefore different families of laws. This does not imply arbitrariness. Models must compress the same observational distribution and remain stable under their own implications. The resulting plurality, if it occurs, is constrained plurality.

The aim of the framework is not to replace empirical science or traditional epistemology, but to describe the structural constraints under which an epistemic system embedded within the universe must operate.

The formal architecture precisely locates three problems that resist full resolution from within any closure: the problem of will, the problem of distinguishing genuine from merely apparent epistemic openness, and the problem of the criterion by which a system recognises new observations as demanding refinement. The paper argues that locating these problems with formal precision is itself a contribution—that a framework which shows exactly where explanation runs out is preferable to one that conceals those limits behind descriptive fluency.

2 Relation to Existing Approaches

The framework developed here sits at the intersection of several existing lines of research, while differing from each in its formal treatment of embeddedness, representational closure, and model-relative structure.

Most directly, it formalises central commitments of Hilary Lawson's closure theory. Lawson argues that the world as encountered is always a world fixed by closure, that openness underlies and escapes every closure, and that the question of which closures to adopt is therefore irreducibly evaluative (Lawson, 2001). The present framework gives these claims a precise structural expression: the inference–implication loop is the closure mechanism, $\mathcal{W}^*$ is the space of stable closures, the

quotient space is the closed texture, and the inclusion $C \subseteq D$ is the formal statement that evaluation is interior to the representational structure rather than prior to it. The analysis of institutions and refinement extends this picture by showing that the evaluative dimension of closure is not merely a feature of individual systems but is transmitted, compressed, and potentially lost across generations.

The account also bears comparison with predictive and Bayesian approaches in contemporary philosophy of mind and cognitive science. Predictive processing models treat cognition as the continuous generation of predictions that are compared with incoming sensory signals, with discrepancies driving model revision (Clark, 2016; Friston, 2010). The inference–implication loop introduced here has a related structure: observations generate models through the inference map F, while models generate observational implications through the map g. However, the present framework differs from predictive-processing accounts in one crucial respect: both observations and models are treated as structures internal to a single universe rather than as elements of an external inference problem. The framework therefore addresses not only how models are updated, but how coherence is to be defined for an epistemic system that has no access to an external vantage point.

In philosophy of science, the view developed here is also close in spirit to structural realism. Structural realists argue that scientific knowledge concerns the relational structure of the world rather than the intrinsic nature of unobservable entities (Worrall, 1989; Ladyman, 1998). In the present framework, relational structure appears in an explicitly model-relative mathematical form. Each self-consistent world model w induces a classifier $\pi_w : D \to Z_w$ that partitions the observation space, and empirical regularities arise as relational invariants in the quotient space $Q_w = D/\!\sim_w$ determined by that partition. What embedded observers identify as physical laws are therefore relational structures within a representational quotient induced by the model. In this sense the framework provides a formal account of how structural knowledge arises from representational compression.

This model-relative account of law also bears comparison with relational approaches in physics. Rovelli's relational quantum mechanics emphasises that physical properties are defined relative to interactions rather than to absolute external states (Rovelli, 1996). Physical laws in the present framework are likewise relational invariants, though the present argument grounds their model-relativity in epistemological rather than specifically physical considerations.

The entropy measure introduced here connects the framework to statistical mechanics and information theory. Shannon introduced entropy as a logarithmic measure of expected surprisal associated with a probability distribution (Shannon, 1948). Jaynes later interpreted statistical mechanics as inference over probability distributions subject to informational constraints (Jaynes, 1957). The present framework recovers entropy as a consequence of representational compression rather than positing it as primitive: the classifier π_w partitions the observation space into equivalence classes, and the entropy $H(w)$ measures the expected surprisal of those classes. The framework does not claim identity between this quantity and thermodynamic entropy; rather, it argues for a structural convergence between them, grounded in their shared dependence on the partitioning of a probability space.

In biology and systems theory, Maturana and Varela described cognition as arising from operational closure within self-referential systems (Maturana and Varela, 1980; Varela et al., 1991). The self-consistency condition $T(w) = w$ is a formal analogue of operational closure, with the additional feature that the closed system contains its own evaluative structure as classified content. Read in the present terms, closure is reproduced not merely from isolated outputs but from the structured observational profile a model makes possible from within. The dynamical structure of agent–environment interaction underlying such representational frameworks is analyzed in *The Imagination Machine IV: Systems*.

Finally, the social extension of the framework places it in conversation with social epistemology.

Longino and Kitcher have both argued, in different ways, that knowledge is constitutively social and that the norms governing inquiry are sustained and revised by communities rather than by isolated individuals (Longino, 1990; Kitcher, 1993). The institutional analysis developed here is consistent with this emphasis while grounding it in the formal architecture of the framework. The distinction between generative and compressed inheritance corresponds, at the social level, to the difference between communities that transmit the capacity for inquiry and communities that merely conserve its prior outputs.

3 The Block Universe and the Derivation of $C \subseteq D$

Let Ω denote the universe. Define the following subsets:

$$D \subseteq \Omega \quad \text{(the set of observations)}$$

$$\mathcal{W} \subseteq \Omega \quad \text{(the set of world models)}$$

$$C \subseteq \Omega \quad \text{(the set of classifiers).}$$

We argue for, rather than merely stipulate, the inclusion

$$C \subseteq D \subseteq \Omega.$$

The argument proceeds from the conditions of self-aware representation. Consider what distinguishes an epistemic system—a genuine subject—from a mere transducer. A thermostat classifies temperature, but its classification is not available to it as an object of experience. It cannot encounter its own sorting activity as something that could have been otherwise. An epistemic system, by contrast, is one whose classificatory acts are themselves accessible to it: it can attend to how it is attending, sort its ways of sorting, and in principle revise the dispositions that govern its encounter with the world.

This reflexive accessibility is not an optional feature added to an otherwise complete epistemic system. It is the condition that makes a system epistemic in the first place. A system that cannot encounter its own classifiers cannot recognise itself as one possible closure among others, cannot doubt its own representations, and therefore cannot be said to know in any sense that involves the distinction between appearance and reality. Cartesian doubt is only possible for a system whose classificatory acts are elements of its observation space.

The inclusion $C \subseteq D$ is therefore a transcendental condition: any system that satisfies the minimal criterion for being an epistemic subject must satisfy it. The formal apparatus of this paper applies precisely to systems meeting that criterion.

Remark 1 (Reflexivity Without Vicious Regress)**.** *The condition $C \subseteq D$ means that the system can classify its own classifiers. One might worry that classifying a classifier requires a further classifier, which requires a further classifier still, generating an infinite regress. This regress does not arise in the block universe framing because that framing is atemporal: all observations, including observations of classifiers, are simultaneous elements of the single relational structure Ω. The self-consistency condition $T(w) = w$, developed in Section 10, is a fixed-point condition rather than a termination condition. What matters is not that the regress terminates in a foundation but that the loop closes on a stable fixed point.*

4 World Models and Classification

Each world model $w \in \mathcal{W}$ induces a classifier

$$\pi_w : D \to Z_w$$

where $Z_w \subseteq D$. Thus a model compresses observations by mapping them to representative observational states. Because $C \subseteq D$, the domain of π_w includes classifiers themselves. A world model therefore classifies not only raw observational content but also the evaluative and selective dispositions of the system that holds it.

Remark 2 (Representational Witness)**.** *The condition $Z_w \subseteq D$ ensures that every abstract class induced by π_w is instantiated by at least one observational state. The representative is not assumed to be unique or privileged; it merely witnesses the existence of the class.*

Definition 1 (Model-Induced Equivalence Relation)**.** *For $d_1, d_2 \in D$ define*

$$d_1 \sim_w d_2 \quad \textit{iff} \quad \pi_w(d_1) = \pi_w(d_2).$$

Definition 2 (Equivalence Class)**.** *For $d \in D$, define*

$$[d]_w = \{d' \in D \mid \pi_w(d') = \pi_w(d)\}.$$

The classifier therefore induces a partition of the observation space. When d is itself a classifier—that is, when $d \in C$—its equivalence class $[d]_w$ groups together all observational states that the world model treats as equivalent ways of sorting the world. Different valuations may thus collapse into the same equivalence class under a given model, or be distinguished by a more refined one.

5 Valuation and Will as Interior Observations

The inclusion $C \subseteq D$ has a consequence that deserves explicit statement before the formal development continues.

Valuation—the assignment of significance to observations—and will—the selective pressure that drives a system toward one closure rather than another—are traditionally treated as standing outside epistemological frameworks. They appear as boundary conditions: given that a system values certain outcomes, what can it know? The present framework does not dissolve this exterior status so much as restate it with formal precision.

If the acts by which a system evaluates and selects are themselves classifiers, and if classifiers are observations, then valuation and will are elements of D. They are subject to the same measure μ_D, the same quotient structure induced by π_w, and the same representational compression as any other observation. A self-consistent world model does not merely organise perceptual content; it also classifies the evaluative structure through which the system engages the world.

This does not reduce will to mechanism, nor does it claim to resolve the problem of agency. What it establishes is more modest and more precise: will appears within D, is partially compressed by every model, and yet is not exhausted by any compression. This is not because will is supernatural or causally unconstrained, but because it is the condition under which the world becomes held as anything at all—the potentiality that precedes and exceeds any particular representation of it. The formal loop determines the space of stable closures $\mathcal{W}^*$, but the selection of a particular element from that space is precisely what the framework locates as irreducible. Willing is not explained

away; it is what remains when the inference–implication loop has done everything it can do—not a gap in the framework, but the condition the framework must include without being able to absorb.

Metaphysical closure is therefore prevented not by any deficiency of the representational apparatus, but by what the apparatus must include: the very acts of valuation that animate it. The framework's contribution here is not resolution but precision—knowing exactly where the limit lies is different from not knowing where to look.

6 Statistical Structure

Assume the observation space carries a probability structure

$$(D, \Sigma_D, \mu_D)$$

where Σ_D is a σ-algebra and μ_D a probability measure.

The measure μ_D is the principal way in which observational structure constrains closure. It prevents the framework from collapsing into the view that any self-supporting classificatory system is epistemically on a par with any other. Models partition one and the same observational space, and the measure of those partitions is not up to the model alone.

Proposition 1 (Measurable Partition)**.** *If each π_w is measurable and Z_w carries a σ-algebra in which singletons are measurable, then the equivalence classes $[d]_w$ form a measurable partition of (D, Σ_D, μ_D).*

Proof. Since π_w is measurable, the preimage of each singleton in Z_w lies in Σ_D. But

$$[d]_w = \pi_w^{-1}(\{\pi_w(d)\}),$$

so each equivalence class is measurable. The classes partition D by construction. □

Lemma 1 (Probability of Classes)**.** *For any model w,*

$$\sum_{[d]_w \in Q_w} \mu_D([d]_w) = 1.$$

Proof. The sets $[d]_w$ form a measurable partition of D. Since μ_D is a probability measure on D, the total measure of the partition equals $\mu_D(D) = 1$. □

Remark 3 (Origin and Calibration of the Observational Measure)**.** *The probability measure μ_D represents the empirical distribution of observations across the observation space D. Conceptually it may be understood in several compatible ways.*

First, it may represent the long-run frequency distribution of observations generated across the ensemble of observers embedded in Ω. Since D contains the observations of all observers, the measure aggregates the empirical structure encountered throughout the block universe. This need not be understood as arbitrary sampling from an undifferentiated flux. In many natural settings, observers are embedded in environments structured by stable but incommensurate dynamical cycles whose relative phases continually drift without exact repetition. Under such conditions, sequential observation repeatedly samples a structured signal that is neither perfectly periodic nor wholly unconstrained. The result is an empirical distribution over observational states: enough recurrence for stable frequencies to emerge, enough phase drift for novelty to persist. On this view, μ_D arises from the statistical structure induced by the dynamical environment in which embedded observers occur.

Second, μ_D may be interpreted inferentially. Following the information-theoretic programme associated with Jaynes, probability distributions can be understood as representations of incomplete knowledge subject to constraints. Under this interpretation μ_D encodes the informational constraints under which an embedded epistemic system performs inference.

These two readings are compatible. A structured observational environment gives rise to stable empirical frequencies, while inference treats those frequencies as constraints on admissible closure. The framework therefore does not require commitment to probability as either purely objective or purely epistemic. What matters structurally is that all world models compress the same observational distribution. This shared measure prevents the space of self-consistent closures from collapsing into arbitrary coherent systems.

However, the compatibility of these two readings is itself a condition that can be satisfied or failed. Call this condition calibration*: the alignment between a system's inferential μ_D—the weights it brings to inference—and the actual empirical distribution of observations in its environment. Calibration is an achievement rather than a default. It can fail in at least two ways. A system may be* miscalibrated*: its inferential weights systematically diverge from actual observational frequencies, producing self-consistent closures that are stable relative to the wrong measure. Such a system refines willingly and generates genuine laws—but laws of a distribution that does not reflect the environment it inhabits. Miscalibration is therefore distinct from both dogmatism and ordinary error: the closure is open to refinement, yet refinement proceeds against a distorted image of the world. Calibration can also fail under* distributional shift*: in genuinely novel environments, a system's inferential μ_D is an extrapolation from past frequencies into regions where those frequencies no longer apply. The alignment between the two readings breaks down precisely where epistemic pressure is greatest.*

Miscalibration thus constitutes a third structural location of epistemic risk, alongside dogmatic refusal to refine and the irreducible remainder of will. The framework diagnoses all three as failures at different levels of the hierarchy $(F, g) \to T \to \mathcal{W}^* \to \pi_w \to Q_w \to R_w$: dogmatism is a failure at the level of (F, g); miscalibration is a failure at the level of μ_D itself, prior to the construction of any particular closure; and will names the underdetermination that persists even when both are functioning well.

7 Representational Quotient

Each model induces a quotient space

$$Q_w = D/\!\sim_w .$$

The elements of Q_w represent observational states modulo the classification performed by the model. This is the closed texture through which the world is encountered: not the world as it is prior to closure, but the world as fixed by the representational intervention of π_w.

Because $C \subseteq D$, the quotient space Q_w contains equivalence classes of classifiers alongside equivalence classes of other observations. The closed texture therefore includes, within itself, the compressed image of the evaluative structure of the system that produced it.

To collect these model-relative quotient spaces into a single ambient codomain, define

$$Q := \bigsqcup_{w \in \mathcal{W}} Q_w,$$

the disjoint union of all quotient spaces induced by world models in $\mathcal{W}$. Thus each Q_w is canonically embedded in Q, while remaining distinguished from $Q_{w'}$ when $w \neq w'$.

8 Implication

For each model $w \in \mathcal{W}$, let

$$\Gamma_w$$

denote the set of canonical observational profiles induced by w, where each such profile is structured in the quotient space Q_w. These profiles are not single isolated observations, but model-relative patterns of observational life: structured ways the world becomes legible from within the closure determined by w.

Define the ambient profile space

$$\Gamma := \bigsqcup_{w \in \mathcal{W}} \Gamma_w.$$

World models produce canonical observational profiles through a map

$$g : \mathcal{W} \to \Gamma$$

such that, for each model $w \in \mathcal{W}$,

$$g(w) \in \Gamma_w \subseteq \Gamma.$$

Thus g assigns to each world model a model-relative observational profile internal to the closure induced by that very model.

9 Inference

Canonical observational profiles generate world models through

$$F : \Gamma \to \mathcal{W}.$$

10 The Consistency Loop

The system is governed by the pair of maps

$$\Gamma \xrightarrow{F} \mathcal{W} \xrightarrow{g} \Gamma.$$

Define the induced operator

$$T = F \circ g : \mathcal{W} \to \mathcal{W}.$$

Definition 3 (Self-Consistent World Model)**.** *A model w is self-consistent if $T(w) = w$.*

Define

$$\mathcal{W}^* = \{w \in \mathcal{W} \mid T(w) = w\}.$$

Self-consistent models reproduce themselves when inference is applied to their own implied observational profiles. In Lawson's terms, they are stable closures: the system's representational intervention reproduces itself under the loop of implication and re-inference. More precisely, a self-consistent model is one whose implied observational profile, when re-submitted to inference, regenerates the same model.

A natural worry is that the fixed-point condition may be too weak: if the maps F and g are unconstrained, perhaps trivial fixed points proliferate. That worry is legitimate in the abstract. The framework does not claim that every fixed point is equally significant. Its claim is that any epistemically admissible closure must at least satisfy this condition, and that the observational measure μ_D together with the refinement structure developed below provides a basis for distinguishing empty stability from informative stability.

Remark 4 (Existence of Fixed Points)**.** *The framework defines epistemically admissible closures as fixed points of the operator $T = F \circ g$. The formal development does not assume that fixed points exist for arbitrary choices of F and g. Rather, the framework identifies a structural condition that any stable closure must satisfy if it exists.*

In many natural settings fixed points arise under mild assumptions. For example, if $\mathcal{W}$ is endowed with a compact topology and T is continuous, Schauder's fixed-point theorem ensures the existence of at least one $w^* \in \mathcal{W}$ such that $T(w^*) = w^*$.

In algorithmic or statistical settings the operator may instead be interpreted as an iterative update rule whose empirical convergence defines the effective closure.

The present framework therefore does not claim that all conceivable inference– implication structures admit stable closures. It instead provides the formal characterisation that any such closure must satisfy when it occurs. In this sense the framework is generative: it specifies meta-structural constraints that a world model must satisfy in order to reproduce itself under the inference–implication loop.

Remark 5 (Plurality of Stable Closures)**.** *Nothing in the framework requires $\mathcal{W}^*$ to be a singleton. Multiple incompatible self-consistent models may coexist as elements of $\mathcal{W}^*$. This plurality is not a defect. It corresponds directly to Lawson's insistence that no single closure is metaphysically privileged. The operator T determines the space of possible stable closures, but it does not determine which element of $\mathcal{W}^*$ is instantiated.*

11 Relational Structure

For each integer $i \geq 1$ define

$$K_i(Q_w) = Q_w^i,$$

the i-fold Cartesian product of Q_w with itself. Thus an element of $K_i(Q_w)$ is an ordered tuple

$$\tau = ([d_1]_w, [d_2]_w, \ldots, [d_i]_w).$$

Let

$$K(Q_w) = \bigsqcup_{i=1}^{\infty} K_i(Q_w) = \bigsqcup_{i=1}^{\infty} Q_w^i,$$

the disjoint union of all finite Cartesian powers of Q_w, collecting relational tuples of every arity into a single set.

Elements of $K(Q_w)$ represent finite relational configurations among equivalence classes of observations, together with their arities. A relational classifier

$$R_w : K(Q_w) \to Q_w$$

assigns canonical relational consequences within the quotient space.

12 Physical Law

Definition 4 (Relational Equivalence)**.** *For $\tau_1, \tau_2 \in K(Q_w)$ define*

$$\tau_1 \sim_{R_w} \tau_2 \quad \textit{iff} \quad R_w(\tau_1) = R_w(\tau_2).$$

Definition 5 (Physical Law)**.** *A physical law under a model w is a relational equivalence class*

$$L = [\tau]_{R_w}$$

for some $\tau \in K(Q_w)$.

Physical laws appear as relational structures within the quotient representation induced by a self-consistent world model. They are stable patterns in the closed texture, not features of an independently accessible world. Different elements of $\mathcal{W}^*$ may induce different quotient spaces and therefore different relational invariants; which laws appear depends on which closure is sustained.

This model-relativity should not be confused with arbitrariness. Any such law is still a law of one and the same observational world as compressed under a particular stable closure. If multiple closures persist, they persist under the constraint of the same D and the same μ_D.

13 Entropy

The classifier π_w compresses the observation space. In this section we assume that the partition $Q_w = D/\sim_w$ is finite or countable, so that the sums below are well defined.

Definition 6 (Class Measure)**.**

$$M_w(d) = \mu_D([d]_w).$$

Definition 7 (Model-Relative Surprisal)**.**

$$S_w([d]_w) = -\log \mu_D([d]_w).$$

Definition 8 (Model-Relative Entropy)**.**

$$H(w) = -\sum_{[d]_w \in Q_w} \mu_D([d]_w) \log \mu_D([d]_w).$$

The quantity $S_w([d]_w)$ measures the probability mass of the equivalence class $[d]_w$, which is the fiber of the projection $\pi_w : D \to Q_w$. The quantity $H(w)$ is the expected surprisal induced by the partition defined by π_w and therefore measures the representational compression associated with the model.

Because classifiers are elements of D, both surprisal and entropy assign measure-theoretic weight not only to equivalence classes of perceptual content but also to equivalence classes of valuations. A valuation that is rare in D carries high surprisal. A coarse model that collapses many distinct valuations into a single class yields low surprisal for that class and lowers the effective distinguishability of evaluative structure. Entropy is therefore not merely a feature of perceptual content; it also measures the coarseness with which a model distinguishes the system's evaluative dispositions.

A note on scope is warranted. The entropy $H(w)$ defined above is a Shannon-type quantity derived from representational compression. The framework does not claim identity between this quantity and thermodynamic entropy. It claims structural convergence: both quantities arise from the same underlying operation of partitioning a probability space, and Jaynes' programme of deriving statistical mechanics from inference over probability distributions subject to informational constraints suggests that this convergence is not superficial. The precise conditions under which model-relative entropy and thermodynamic entropy coincide are left for subsequent work.

14 Representational Refinement

Definition 9 (Refinement)**.** *A model w_2 refines w_1 if $[d]_{w_2} \subseteq [d]_{w_1}$ for all $d \in D$.*

Theorem 1 (Monotonicity of Surprisal)**.** *If w_2 refines w_1, then*

$$S_{w_2}([d]_{w_2}) \geq S_{w_1}([d]_{w_1}).$$

Proof. Refinement implies $[d]_{w_2} \subseteq [d]_{w_1}$, so $\mu_D([d]_{w_2}) \leq \mu_D([d]_{w_1})$. Applying $-\log$ reverses the inequality. □

Theorem 2 (Entropy Equality for Equivalent Observations)**.** *If $d_1 \sim_w d_2$, then $S_w([d_1]_w) = S_w([d_2]_w)$.*

Proof. If $d_1 \sim_w d_2$ then $[d_1]_w = [d_2]_w$, so $\mu_D([d_1]_w) = \mu_D([d_2]_w)$ and the definition of S_w yields the result. □

14.1 Refinement Drives Evolution

It is a common intuition that refinement—the transition from w_1 to w_2 where $[d]_{w_2} \subseteq [d]_{w_1}$—represents a "narrowing in" on a point-like truth. However, the framework suggests something orthogonal. As the partition becomes finer, the measure μ_D associated with each class decreases, and the expected surprisal increases. The more finely one's categories partition implied observational profiles–the more specific one's prediction becomes–the more likely it becomes that observation, when resubmitted to inference, will evolve one's world model.

15 Interpretation

The hierarchy of structure is

$$(F, g) \to T \to \mathcal{W}^* \to \pi_w \to Q_w \to R_w.$$

The condition $C \subseteq D$ runs through every level of this hierarchy. Classifiers enter the observation space as observations, are compressed by π_w, appear in the quotient space Q_w, figure in relational tuples in $K(Q_w)$, and carry surprisal under S_w. Valuation is not a parameter set from outside the system; it is a structural feature of the observation space that the system's own representational apparatus must absorb, compress, and partially lose.

The implication map now makes explicit that closure is reproduced not from an atomized observational residue but from a structured observational profile. A stable closure is therefore a view from somewhere in the strict sense: a model whose own internally generated way of inhabiting the world, when reinterpreted through inference, yields the same model again.

16 Institutions as Intergenerational Compression

The framework developed so far treats an epistemic system as a single relational structure. But embedded systems are not isolated. They exist within communities of systems that share, contest, and transmit closures across time. This section extends the framework to that social dimension, focusing on the role of institutions.

The central observation is this: no individual knower transmits a closure to a successor by reproducing the full observation space D that gave rise to it. What is transmitted is always a

compression—a residue of the inferential work that produced a given $w^* \in \mathcal{W}^*$. Institutions are the mechanisms by which this intergenerational compression is stabilised.

More precisely, what passes between generations is not the loop itself—the maps F and g that generated the fixed point—but a projection of the implied observational profile and the quotient structure it presupposes into the observation space of the successor generation. The successor receives the closed texture without necessarily receiving the closure mechanism. Institutions are the structures within Ω that perform and stabilise this projection, re-embedding the inherited profile of closure as observations in the successor's D, making it available for classification by the successor's own π_w.

This framing carries an immediate consequence. A successor generation may inherit a stable closure without inheriting the capacity to regenerate it under pressure from new observations. The quotient structure arrives, but the inferential machinery that produced it does not.

We distinguish two modes of institutional transmission. *Compressed inheritance* transmits the closed profile alone: the successor can apply the inherited partition but cannot update it. *Generative inheritance* transmits F and g alongside that profile: the successor can regenerate the closure from within, extend it, and revise it when new observations demand a finer partition.

The distinction matters because the observation space D does not stand still. New observations enter D in every generation, and a partition that was self-consistent under an earlier μ_D may fail to remain so as the measure shifts. A generatively inherited closure can meet this pressure; a compressedly inherited one cannot. The institution that transmits only the quotient structure is therefore more fragile—not because it contains false beliefs, but because it has lost the capacity to refine.

Note that institutions may also transmit miscalibrated measures. A community that inherits both F and g alongside a systematically distorted μ_D possesses the machinery for refinement while lacking accurate observational weights on which to exercise it. Generative inheritance is therefore necessary but not sufficient for epistemic health: the inferential measure must also track the environment it purports to represent.

17 Knowledge, Dogma, and the Structure of Refinement

A natural question arises from the plurality of stable closures established in Section 10: if $\mathcal{W}^*$ may contain many incompatible elements, and the framework provides no external criterion for preferring one over another, how does it distinguish knowledge from dogma? Both are self-consistent. Both survive the inference–implication loop. Both can be institutionally transmitted.

The answer is that the distinction does not require an external criterion. It falls out of the structure already in place, specifically from the relationship between a closure and its behaviour under refinement.

Recall that a model w_2 refines w_1 when $[d]_{w_2} \subseteq [d]_{w_1}$ for all $d \in D$. Refinement always costs higher surprisal: a finer partition assigns lower probability mass to each class and therefore higher S_w to each observation. A closure disposed toward knowledge is one that remains willing to pay this cost—one whose inference–implication loop, when supplied with observations that increase the consistency gap under the current partition, responds by generating a finer π_w rather than forcing the new observations into existing classes.

Dogmatic closure is precisely the refusal to pay this cost. A dogmatic model maintains its self-consistency not by genuinely absorbing new observations but by compressing them into existing equivalence classes regardless of their character. New elements of D are mapped by π_w to existing elements of Z_w even when a more faithful compression would require extending Z_w. The partition

is held fixed; the observations are bent to fit it.

Miscalibration, introduced in Remark 3, constitutes a distinct failure mode. A miscalibrated closure may be fully open to refinement—willing to extend Z_w whenever the consistency gap demands it—and yet refine systematically against a distorted image of the observational world. Where dogmatism is a failure of disposition at the level of (F, g), miscalibration is a failure of the measure μ_D itself, prior to any particular act of closure. The two failures are formally separable: a closure can be dogmatic without being miscalibrated, or miscalibrated without being dogmatic, or both simultaneously.

A clarification is required here. The criterion just stated relies on a notion of stable absorption that is not itself fully decidable from within a single closure. Determining whether a new observation d genuinely strains the existing partition or is legitimately compressed into it requires assessing the consistency gap, and different closures may assess that gap differently. The framework does not resolve this from outside; it rather establishes the vocabulary within which the question can be precisely posed and contested. The distinction between knowledge and dogma is therefore best understood as identifying a structural disposition—the preparedness to extend Z_w under pressure—rather than as a decision procedure that can be applied mechanically from within any single closure. Crucially, this question is available to any system satisfying $C \subseteq D$, since such a system can observe its own classificatory behaviour and the consistency of its loop.

Several further consequences follow. First, the distinction is not binary but gradational. A closure may be refinable with respect to some regions of D while dogmatic with respect to others. Institutions that transmit F and g alongside the inherited profile of closure preserve the capacity for refinement, but may do so selectively—maintaining the inferential machinery for some domains while suppressing it for others.

Second, the surprisal cost of refinement explains a persistent feature of actual epistemic communities. Dogmatic compression avoids this cost by refusing to see new observations as genuinely new. Coarser models assign lower surprisal to the observations they assimilate, and lower surprisal feels, from within the closure, like greater understanding. The framework thus provides a structural account of why the pressure toward dogmatic closure is not merely psychological but has a measure-theoretic basis.

Third, because $C \subseteq D$, the distinction applies to evaluative structure as well as perceptual content. A closure that refuses to refine its classification of classifiers—that compresses distinct valuations into the same equivalence class regardless of the observational pressure to distinguish them—is dogmatic about value in precisely the same structural sense. The framework does not treat these as different in kind.

Returning to the hierarchy established in Section 15, the distinction between knowledge and dogma lives at the level of (F, g) rather than at the level of $\mathcal{W}^*$. Two closures may be indistinguishable as fixed points—equally self-consistent, equally stable—while differing fundamentally in whether the loop they instantiate remains open to refinement. Stability is not the same as openness, and it is openness to refinement—the disposition to pay the surprisal cost when the consistency loop demands it—that the present framework identifies as the structural mark of what distinguishes knowledge from its appearance.

18 Conclusion

Embedded epistemic systems cannot appeal to external correspondence as their standard of coherence. Coherence appears instead as internal closure of the inference–implication loop under the statistical structure of observations. Self-consistent world models arise as fixed points of the

operator this loop induces, and each such model compresses the observation space into a quotient representation whose relational invariants constitute physical law and whose measure-theoretic multiplicity constitutes entropy.

The structural feature that distinguishes this framework from earlier accounts is the inclusion $C \subseteq D$: classifiers are observations. This inclusion is not stipulated but derived—it is the transcendental condition on any system capable of Cartesian doubt, any system that can recognise itself as one possible closure among others. This means that valuation and will—the dispositions that select among possible closures—are interior to the representational architecture. They appear in the observation space, are subject to compression, and leave their trace in the quotient structure. Yet they are not exhausted by any compression. The formal loop determines the space of stable closures, but not which closure is instantiated. This remainder is not a gap in the framework; it is the constitutive openness that the inference–implication loop must encompass but cannot exhaust.

The implication map clarifies the form of that closure. What a model implies is not merely an isolated consequence but a canonical observational profile internal to the model itself: a structured way the world appears from somewhere. A self-consistent world model is therefore one whose own implied profile of observational life, when reinterpreted through inference, reproduces that same model. Stable theory and stable world-profile co-arise.

The social extension of the framework yields two further results that follow from the same architecture without requiring external normative imports. Institutions are the mechanisms by which stable closures are transmitted across generations, but they transmit closures in two structurally distinct modes: generative inheritance conveys the inferential machinery alongside the fixed point, while compressed inheritance conveys only the inherited profile of closure. And the distinction between knowledge and dogma reduces, within the framework, to the distinction between closures that remain open to refinement and those that hold their partition fixed against the pressure of new observations—a difference that identifies a structural disposition rather than a decision procedure applicable from outside any particular closure.

The framework thus diagnoses three irreducible structural locations of epistemic risk. Dogmatism is a failure of disposition at the level of (F, g): the loop exists but refuses to refine. Miscalibration is a failure at the level of μ_D: the loop refines willingly but against a distorted image of the world. And will names the underdetermination that persists even when both are functioning well—the necessity of choosing a closure in territory no model can fully exhaust. Together these three constitute the complete formal topology of epistemic failure for an embedded system.

Epistemic closure, physical law, entropy, and the social conditions of knowledge therefore emerge as successive consequences of a single embedded representational architecture. What prevents metaphysical closure—what keeps the system in relation to the openness beneath its representations—is the evaluative structure that the architecture must include but cannot fully exhaust.

References

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

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

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, 106(4):620–630.

Kitcher, P. (1993). *The Advancement of Science: Science without Legend, Objectivity without Illusions*. Oxford University Press.

Ladyman, J. (1998). What is structural realism? *Studies in History and Philosophy of Science Part A*, 29(3):409–424.

Lawson, H. (2001). *Closure: A Story of Everything*. Routledge.

Longino, H. E. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

Maturana, H. R. and Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel.

Rovelli, C. (1996). Relational quantum mechanics. *International Journal of Theoretical Physics*, 35(8):1637–1678.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Varela, F. J., Thompson, E., and Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press.

Worrall, J. (1989). Structural realism: The best of both worlds? *Dialectica*, 43(1–2):99–124.

The Imagination Machine II: Relational Invariants, Quotient Structure, and the Reproducibility of Science

Mark Tracy
Boston University
mrktracy@bu.edu

Abstract

Scientific knowledge stabilizes through the reproducibility of experimental results across independent observers and experimental contexts. This paper interprets reproducibility through the compression–extension architecture developed in the Imagination Machine series. Observational data are first produced in highly indexical form, tied to particular observers, instruments, and experimental circumstances. Scientific modeling compresses these observations through a classifier that quotients away observational detail while preserving selected relational invariants. A scientific law is then interpreted as a relational structure that remains invariant under this quotient map. Reproducibility corresponds to the stability of these invariants across independent experiments. From this perspective the methodology of science may be understood as the collective construction of quotient representations of the observational world, within which invariant relations appear as physical law.

1 Introduction

The Imagination Machine series develops a formal framework for embedded epistemic systems. In this framework an agent constructs a world model by iteratively compressing observational data into a representation that preserves relational structure while discarding irrelevant detail. The admissible models of the system appear as fixed points of the inference–implication loop introduced in the first paper of the series.

A central question in the philosophy of science concerns the reproducibility of experimental results. Independent laboratories performing the same experiment under different conditions frequently obtain observational data that differ in numerous superficial ways. Nevertheless, scientific laws appear as stable regularities that persist across these differences.

The present paper interprets reproducibility as a consequence of the quotient structure induced by representational compression. Scientific laws correspond to relational invariants that remain stable under the quotient map from observational data to scientific representation.

2 Observational Surfaces

Every experiment produces data in a highly indexical form. Observations are tied to particular observers, instruments, experimental procedures, and environmental circumstances.

Definition 1 (Observation Event)**.** *An observation event is a tuple*

$$x = (o, a, t, \ell, p, m)$$

where o denotes the observer, a the apparatus configuration, t the time of observation, ℓ the spatial location, p the experimental protocol, and m the measured outcome.

Let D denote the space of such observation events. Two observation events may differ in many of these parameters while nevertheless expressing the same underlying regularity.

3 Representational Compression

A scientific model compresses the observational surface by mapping observation events into a representation that preserves selected relational structure.

Definition 2 (Scientific Classifier)**.** *Let*

$$\pi : D \to Z$$

be a classifier mapping observation events into representational states Z. The map π induces an equivalence relation on D defined by

$$x \sim_\pi y \quad \text{if and only if} \quad \pi(x) = \pi(y).$$

The quotient space

$$Q = D/\!\sim_\pi$$

groups together observation events that are treated as equivalent by the scientific model.

Remark 1. *The classifier π may include transformations such as coordinate normalization, calibration correction, statistical averaging, or parameter estimation. These operations discard observational detail while preserving relational structure relevant to the theory.*

4 Relational Invariants

Scientific laws correspond to relations that remain invariant across equivalence classes in the quotient representation.

Definition 3 (Relational Invariant)**.** *A relation R defined on the representational space Z is a relational invariant if it holds for all representatives of an equivalence class in Q.*

Examples include the constancy of gravitational acceleration in Newtonian mechanics, the Lorentz invariance of spacetime intervals in relativity, and the ideal gas relation in thermodynamics.

Remark 2. *The invariance of these relations reflects the fact that the observational differences removed by the quotient map do not alter the relational structure preserved by the model.*

5 Reproducibility

The reproducibility of scientific results can now be interpreted as stability under the quotient map.

Definition 4 (Reproducible Result)**.** *An experimental result is reproducible if observation events from independent experiments fall into the same equivalence class of Q under the classifier π.*

In practice this means that while raw measurements may vary across laboratories, the representational compression applied by the scientific model maps them to the same relational structure.

Remark 3. *Experimental methodology exists largely to ensure that independent investigators apply compatible compression maps. Standardized protocols, calibration procedures, and statistical analysis all serve to align the quotient representations used by different laboratories.*

6 Scientific Method as Quotient Construction

The methodology of science may therefore be interpreted as a collective process for constructing quotient representations of observational reality.

Different laboratories act as independent epistemic agents observing the same environment through distinct observational surfaces. A scientific theory stabilizes when the compression map used by these agents yields consistent relational invariants across their respective data.

Proposition 1. *Scientific consensus emerges when independently observed data sets share a common quotient representation under a shared classifier.*

7 Symmetry and Physical Law

Modern physics frequently formulates laws in terms of symmetry principles. These symmetries express invariance under transformations such as spatial translation, temporal translation, or coordinate change.

Within the present framework these symmetries appear naturally as transformations that leave the quotient representation unchanged. A symmetry therefore corresponds to an operation on observation events that preserves equivalence classes in the quotient space.

Remark 4. *This perspective explains the centrality of symmetry in modern physics: symmetry transformations are precisely those operations that preserve the relational invariants retained by the representational compression.*

8 Conclusion

The Imagination Machine framework interprets knowledge formation as the compression of observational data into representations that preserve relational structure. Scientific laws appear as invariants within the quotient representations produced by this compression.

From this perspective the reproducibility of science is not mysterious. Independent experiments produce different observational details, but once those details are quotiented away by the scientific classifier, the same relational invariants emerge. Reproducibility therefore reflects the stability of these invariants across observational contexts.

Scientific practice can thus be understood as a distributed epistemic process in which many observers collaboratively construct quotient representations of the observational world. Physical law corresponds to the relational structure that remains invariant within those representations.

The Imagination Machine III: Systems

Mark Tracy
Boston University
mrktracy@bu.edu

Introduction

This paper is part of a series titled *The Imagination Machine*. The first paper, *The Imagination Machine I: A View from Somewhere*, develops a formal epistemic framework for embedded observers and introduces the inference–implication loop that defines self-consistent world models. Within that framework, observations, classifiers, and world models all appear as structures internal to the same universe, and epistemic coherence arises as the closure of a representational loop rather than correspondence with an external vantage point.

The present paper develops a complementary layer of the project by introducing a general formalism for systems. Whereas the first paper describes the structure of representational closure for embedded epistemic systems, the present work describes the dynamical coupling between components of such systems, particularly in the case of agent–environment interaction. The goal is to define systems in an extremely general way so that the formalism has maximal expressiveness while making minimal assumptions.

In the first section, we develop a general definition of a system in terms of measurable variables, stochastic processes, and functional relations between inputs and outputs. In the following section we introduce optimization models and relate them to systems through the problem of system identification. The agent–environment framework developed later in the paper provides a general structure for modeling adaptive systems whose outputs influence the environment from which future inputs arise.

1 General System Definition[1]

Define a set of variables that can be measured in practice. By necessity, this will be a countable set of variables, even if the underlying real system of interest has uncountable degrees of freedom. By measuring these variables over a set of time points I with minimal value t_0, we collect *data*. We distinguish between two different proper subsets of variables: *input variables* and *output variables*.

[1]Some language and structure adapted from *Introduction to Discrete Event Systems: Third Edition* by Christos G. Cassandras and Stéphane Lafortune. https://doi.org/10.1007/978-3-030-72274-6

1.1 Input Variables

Without loss of generality, we will assume going forward that there is only one input variable. This is without loss of generality because any countable set of input variables may be represented by a tuple whose components are the simpler variables.

We represent the input variables with a random process. In particular, let $(\Omega_u, \mathcal{F}_u, \mathbb{P}_u)$ denote a probability space. Let U_{in} be the set of possible values the input variable may take.

Then the input process

$$u : \Omega_u \times I \to U_{\text{in}}$$

is measurable as a function from $(\Omega_u \times I, \mathcal{F}_u \otimes \mathcal{F}_I)$ to $(U_{\text{in}}, \mathcal{F}_{\text{in}})$, where $\mathcal{F}_I$ denotes a σ-algebra on the time set I, $\mathcal{F}_{\text{in}}$ denotes a σ-algebra on the input space U_{in}, and where the symbol $\otimes$ denotes the product σ-algebra. For each fixed time $t \in I$, the mapping

$$\omega \mapsto u(\omega, t)$$

is a random variable, and for every measurable set $A \subseteq U_{\text{in}}$ (i.e. $A \in \mathcal{F}_{\text{in}}$), the distribution of $u(t)$ is given by:

$$\mathbb{P}_u(\{\omega \in \Omega_u \mid u(\omega, t) \in A\})$$

We note, crucially, that although we are representing our input variable as a random process, input variables are often chosen to be those that one can deliberately vary over time. In such a case, the input variable may not be stochastic. In general, the input variable $u(t)$ at any time $t \in I$ may be a tuple in a product space of simpler, potentially degenerate random variables.

1.2 Output Variables

Complementary to the input variables are the output variables. Again, we will treat the case of a single output variable, since countably many output variables may be treated as a tuple.

Similarly to the input variable, we can represent the output variable as a random process. As before, let $(\Omega_y, \mathcal{F}_y, \mathbb{P}_y)$ denote a probability space. Let U_{out} be the set of possible values the output variable may take.

Then the output process

$$y : \Omega_y \times I \to U_{\text{out}}$$

is measurable as a function from $(\Omega_y \times I, \mathcal{F}_y \otimes \mathcal{F}_I)$ to $(U_{\text{out}}, \mathcal{F}_{\text{out}})$, where $\mathcal{F}_{\text{out}}$ denotes a σ-algebra on the output space U_{out}, and where the symbol $\otimes$ denotes the product σ-algebra. For each fixed time $t \in I$, the mapping

$$\omega \mapsto y(\omega, t)$$

is a random variable, and for every measurable set $A \subseteq U_{\text{out}}$ (i.e. $A \in \mathcal{F}_{\text{out}}$), the distribution of $y(t)$ is given by:

$$\mathbb{P}_y(\{\omega \in \Omega_y \mid y(\omega, t) \in A\})$$

1.3 Relating Inputs to Outputs

The relation between the input variable and time, and the resulting output variable, is given by a functional g. A functional is a function whose domain is a Cartesian product of one or more sets of functions and zero or more other sets. In this case, the domain of the functional g is the Cartesian product of the set $\mathcal{U}$ of all measurable input processes $u : \Omega_u \times I \to U_{\text{in}}$ and the time set I. The codomain of g is $\mathcal{P}(U_{\text{out}}, \mathcal{F}_{\text{out}})$, the space of probability measures over the measurable space $(U_{\text{out}}, \mathcal{F}_{\text{out}})$. Explicitly:

$$g : \mathcal{U} \times I \to \mathcal{P}(U_{\text{out}}, \mathcal{F}_{\text{out}}),$$

The functional g satisfies:

$$y(t) \sim g[u, t]$$

Or, if modeling time as discrete, where t_{i+1} is a successor of t_i in a countable and strictly ordered time set I whose minimal element is t_0:

$$y(t_{i+1}) \sim g[u, t_i]$$

The symbol $\sim$ denotes random sampling or should be read as "is distributed according to." If the distribution is degenerate (i.e. there is no stochasticity), then the symbol may be treated as deterministic assignment, identically to an "equals" sign. In other words, determinism is represented as stochasticity with a degenerate distribution—assigning probability 1 to a single outcome.

Note that each component of the output variable (when considering a tuple of simpler variables) at time t can depend in general on the value of any component of the input variable (again, when considering a tuple of simpler variables) at any subset of time points, potentially including future points. While many physical systems are assumed to be "causal" (outputs depend only on present and past inputs), the mathematical formulation permits non-causal dependencies, allowing flexibility in modeling retroactive influence.

The functional g relating input and time to output may be an evaluation functional, which directly evaluates an input variable at a given time point, e.g.:

$$y(t) = g[u, t] = u(t)$$

It may also be a function of such evaluation functionals, e.g.,

$$y(t) = g[u, t] = u(t) + 3u(t - 1.3) - 76.8u(t + 4)^2$$

1.4 State

While the above is a general description of any system, in many cases, especially where history and memory matter, we find it useful to model the system's internal condition explicitly.

This internal condition is what we call the system's *state*, which we can represent as a random process defined over some probability space $(\Omega_s, \mathcal{F}_s, \mathbb{P}_s)$. Letting U_{state} be the set of values that the state may take, we can write:

$$s : \Omega_s \times I \to U_{\text{state}}$$

Again, we consider the state $s(t)$ at time t to be a single random variable without loss of generality, since the state variable may be a tuple in a product space of simpler, potentially degenerate (i.e. determinate) random variables.

The evolution of the state may be represented in continuous time by a stochastic differential equation:

$$\dot{s}(t) \sim f[u, s, t], \quad s(t_0) \sim s_0 \quad \text{for some } s_0 \in \mathcal{P}(U_{\text{state}}, \mathcal{F}_{\text{state}})$$

where $\mathcal{F}_{\text{state}}$ is a σ-algebra on U_{state} and where

$$f : \mathcal{U} \times \mathcal{S} \times I \to \mathcal{P}(U_{\text{change}}, \mathcal{F}_{\text{change}}),$$

for some set U_{change} whose elements represent rates of change of the state and for $\mathcal{F}_{\text{change}}$ a σ-algebra on U_{change}; and where we denote by $\mathcal{S}$ the space of all measurable state processes $s : \Omega_s \times I \to U_{\text{state}}$.

If modeling time as discrete rather than continuous, then we may represent state dynamics as an update rule:

$$s(t_{i+1}) - s(t_i) \sim f[u, s, t_{i+1}], \quad s(t_0) \sim s_0 \quad \text{for some } s_0 \in \mathcal{P}(U_{\text{state}}, \mathcal{F}_{\text{state}})$$

where, similarly to before,

$$f : \mathcal{U} \times \mathcal{S} \times I \to \mathcal{P}(U_{\text{change}}, \mathcal{F}_{\text{change}}),$$

for some set U_{change} whose elements represent changes in the state, and where t_{i+1} is a successor of t_i in a countable and strictly ordered time set I whose minimal element is t_0.

The relation between the input variable, the system state, and time, and the resulting output variables may then be expressed as a functional:

$$y(t) \sim g[u, s, t]$$

or, in discrete time, where t_{i+1} is a successor of t_i in a countable and strictly ordered time set I whose minimal element is t_0:

$$y(t_{i+1}) \sim g[u, s, t_i]$$

where

$$g : \mathcal{U} \times \mathcal{S} \times I \to \mathcal{P}(U_{\text{out}}, \mathcal{F}_{\text{out}}).$$

Remark 1.1 (Observable Parameters from Admissible Transformations) *The parameters that can be meaningfully measured about a system's state are determined by how it can transform without changing the system's input–output behavior: admissible transformations constitute a group acting on the state space; this group action induces a quotient via orbits, and observable parameters are precisely functions on the resulting quotient space.*

2 Optimization Models

A functional, like those discussed above, is a special kind of function. An optimization model is a function approximator. An optimization model consists of a triplet $(\mathcal{H}, O, A)$ of:

1. A hypothesis space $\mathcal{H}$ (a set of functions);
2. An objective $O : \mathcal{H} \to \mathbb{R}$ (a functional whose domain is the hypothesis space and whose range is real numbers) which gives some signal as to the quality of the approximation; and
3. An optimization algorithm $A : \mathcal{H} \to \mathcal{H}$ (a rule for moving through the hypothesis space), in general utilizing the objective.

When learning inductively from data (that is, when attempting to move from particular examples to general principles), a few additional objects may be appended to the aforementioned triplet; in particular:

4. A dataset $\mathcal{D}$.
5. A (possibly unknown) random process P from which data points are sampled. In other words, data is collected empirically from the world during a time interval I_D with $d_i \sim P(t_i) \quad \forall d_i \in D$, where data point d_i is collected at time t_i. Note that in cases where data points may be assumed to be identically distributed and drawn independently, this amounts to a single distribution. In *active learning*, the algorithm A interacts with the random process P, influencing the empirical dataset $\mathcal{D}$ used during optimization. In other words, data points are not sampled according to P before the commencement of the algorithm A, but rather, the process of data collection is itself influenced by the optimization algorithm.
6. A dataset $\mathcal{D}_{\text{aug}}$, where for all $d_{\text{aug}} \in \mathcal{D}_{\text{aug}}$, there exists a function f, an integer N, and a tuple of elements $t \in \mathcal{D}^N$ such that $d_{\text{aug}} \sim f(t)$, where $\sim$ denotes, as before, stochastic sampling of the (possibly degenerate) random function f. In other words, every element of $\mathcal{D}_{\text{aug}}$ is a (potentially stochastic) function of elements of $\mathcal{D}$.
7. A random process P_{train} by which elements of $\mathcal{D}_{\text{aug}}$ are drawn by the optimization algorithm. In particular, the algorithm A draws at time t_i an element $d_i \sim P_{\text{train}}(t_i)$, where $P_{\text{train}}(t_i)$ is a distribution over $\mathcal{D}_{\text{aug}}$.

An inductive bias is a constraint on the hypothesis space. By traversing the hypothesis space algorithmically, an optimization model is intended to minimize the objective function and thus obtain a good approximation to the function that truly represents the system of interest.

3 System Identification

System identification is the process of utilizing an optimization model to find an approximation to the true dynamics of a system using measurements of its input and output variables. In particular, it is useful when the internal state of a system is not known or its internal dynamics–the stochastic differential (or difference) equation(s) and initial conditions governing the state's trajectory–are not known.

4 Agents

The agent–environment coupling introduced here provides the dynamical structure within which the representational closures described in *The Imagination Machine I: A View from Somewhere* may arise for embedded epistemic systems. In that framework, stable world models emerge as fixed points of an inference–implication loop defined over observations internal to the same universe. The systems formalism developed here provides a concrete representation of the interacting processes through which such observations and models may be generated.

An agent necessarily exists within and is co-constituted with an environment. An agent-environment pair comprises two systems, an agent A and environment E, which are in interchange (feeding back to one another); as well as an initial input to either the agent or the environment. In particular, A takes as input the output of E, and E takes as input the output of A, with the recursion beginning from some set of initial inputs to either system.

Formally, we may represent the recursive dependency between an agent A and an environment E as follows:

$$\begin{aligned} u^A(t) &= y^E(t) \quad \text{(agent receives environment's output as input)} \\ u^E(t) &= y^A(t) \quad \text{(environment receives agent's output as input)} \end{aligned}$$

where:

- $u^A(t)$ is the input to the agent at time t
- $y^A(t)$ is the agent's output at time t
- $u^E(t)$ is the input to the environment at time t
- $y^E(t)$ is the environment's output at time t

The recursion begins from a set of initial inputs:

$$\begin{aligned} u^E(t_0) &\sim u_0^E \quad \text{for some } u_0^E \in \mathcal{P}(U_{\text{in}}^E, \mathcal{F}_{\text{in}}^E) \quad \text{or} \\ u^A(t_0) &\sim u_0^A \quad \text{for some } u_0^A \in \mathcal{P}(U_{\text{in}}^A, \mathcal{F}_{\text{in}}^A) \end{aligned}$$

and the pair evolves together over time, potentially governed by their own internal state dynamics:

$$\begin{aligned} \dot{s}^A(t) &\sim f^A[u^A, s^A, t], \qquad y^A(t) \sim g^A[u^A, s^A, t] \\ \dot{s}^E(t) &\sim f^E[u^E, s^E, t], \qquad y^E(t) \sim g^E[u^E, s^E, t] \end{aligned}$$

for some functionals defined analogously as before:

$$
\begin{aligned}
f^A &: \mathcal{U}^A \times \mathcal{S}^A \times I \to \mathcal{P}(U^A_{\text{change}}, \mathcal{F}^A_{\text{change}}) \\
g^A &: \mathcal{U}^A \times \mathcal{S}^A \times I \to \mathcal{P}(U^A_{\text{out}}, \mathcal{F}^A_{\text{out}}) \\
f^E &: \mathcal{U}^E \times \mathcal{S}^E \times I \to \mathcal{P}(U^E_{\text{change}}, \mathcal{F}^E_{\text{change}}) \\
g^E &: \mathcal{U}^E \times \mathcal{S}^E \times I \to \mathcal{P}(U^E_{\text{out}}, \mathcal{F}^E_{\text{out}})
\end{aligned}
$$

That is, each functional takes as input:

- a random input process over I,
- a random state process over I,
- and the current time $t \in I$,

and produces either a rate of change of the state (for f) or an output (for g), potentially by sampling randomly from a distribution of outputs.

4.1 Agents in Discrete Time

In many practical applications, especially in reinforcement learning, the agent-environment interaction is modeled in discrete time. This leads to the following slight change in representation:

$$
\begin{aligned}
u^A(t_{i+1}) &= y^E(t_i) \quad \text{(agent receives environment's most recent output as input)} \\
u^E(t_{i+1}) &= y^A(t_i) \quad \text{(environment receives agent's most recent output as input)}
\end{aligned}
$$

where t_{i+1} is the successor of t_i in some ordered set of time points I (in particular, the time points are indexed by $i \in \mathbb{N}_0$), and where

$$
\begin{aligned}
s^A(t_{i+1}) - s^A(t_i) &\sim f^A[u^A, s^A, t_{i+1}], \qquad y^A(t_{i+1}) \sim g^A[u^A, s^A, t_{i+1}] \\
s^E(t_{i+1}) - s^E(t_i) &\sim f^E[u^E, s^E, t_{i+1}], \qquad y^E(t_{i+1}) \sim g^E[u^E, s^E, t_{i+1}]
\end{aligned}
$$

for some functionals defined analogously as before:

$$
\begin{aligned}
f^A &: \mathcal{U}^A \times \mathcal{S}^A \times I \to \mathcal{P}(U^A_{\text{change}}, \mathcal{F}^A_{\text{change}}) \\
g^A &: \mathcal{U}^A \times \mathcal{S}^A \times I \to \mathcal{P}(U^A_{\text{out}}, \mathcal{F}^A_{\text{out}}) \\
f^E &: \mathcal{U}^E \times \mathcal{S}^E \times I \to \mathcal{P}(U^E_{\text{change}}, \mathcal{F}^E_{\text{change}}) \\
g^E &: \mathcal{U}^E \times \mathcal{S}^E \times I \to \mathcal{P}(U^E_{\text{out}}, \mathcal{F}^E_{\text{out}})
\end{aligned}
$$

That is, each functional takes as input:

- a random input process over I,
- a random state process over I,
- and the current time $t \in I$,

and produces either a state update (for f) or an output (for g), potentially by sampling randomly from a distribution over possible outputs.

4.2 Reinforcement Learning

Reinforcement learning is a special case of an optimization model, whereby the objective O depends on the history of interactions between an agent and its environment and where the algorithm A seeks to maximize the expected cumulative reward obtained through the agent and environment's dynamic coupling.

In the context of the present series, such agent–environment optimization dynamics provide a concrete setting in which representational models may be iteratively refined through interaction with structured environments. A minimal predictive example of such refinement is developed in *The Imagination Machine IV: Linear Encoding of Koopman Spectra in Predictive Agents*.

4.3 Holonic Recursion of Agent–Environment Systems

4.3.1 The Holonic Stack

Definition 4.1 (Level-Indexed Agent–Environment System) *Let $\{H_i\}_{i\in\mathbb{Z}}$ be a countable family of systems indexed by integer level i. For each i, H_i is simultaneously:*

(a) an agent *relative to H_{i+1}, which functions as its environment, so that $u^{H_i}(t) = y^{H_{i+1}}(t)$ and $u^{H_{i+1}}(t) = y^{H_i}(t)$ per the agent–environment coupling of §4;*

(b) an environment *relative to H_{i-1}, which functions as its agent, under the same coupling relation one level down.*

Each H_i possesses its own state process s^{H_i}, input process u^{H_i}, output process y^{H_i}, and governing functionals f^{H_i}, g^{H_i}, exactly as defined for A and E in §4.1–4.2, with the superscript H_i replacing A or E throughout.

A holonic stack *is such a family $\{H_i\}$ together with the coupling relations (a)–(b) holding at every level. A* bounded holonic stack *is a stack with a minimal level i_0 (a level possessing no further environment below it, e.g. a level whose dynamics are taken as primitive) or a maximal level i_n (a level possessing no further environment above it), or both.*

Remark 4.2 *This is a direct generalization of the single agent–environment pair (A, E) of §4: that case is the bounded holonic stack with exactly two levels, $i_0 = A$ and $i_1 = E$ (or, in the unbounded reading, E itself admits further structure as $H_2, H_3, \ldots$, and the pair (A, E) is simply the two lowest levels of a longer stack that §4 does not pursue further). Nothing in §4's definitions of u, y, s, f, or g assumed E was unstructured; the holonic stack makes explicit that E may be given the identical formal treatment as A, recursively.*

Remark 4.3 (Holon as Self-Similar Description) *Each H_i in the stack is described by exactly the same formal apparatus — state, input, output, governing functionals — regardless of i. The holonic stack is therefore a single self-similar construction repeated at every level, rather than a hierarchy requiring level-specific machinery. This matches the informal characterization of a holon in §5.3: each H_i is "a whole unto itself" (it possesses its own complete state-space description as an agent) while being "a participant in" H_{i+1} (it is simultaneously the output-generating component that H_{i+1} receives as input).*

4.3.2 The Three Holonic Channels

The recursive structure of the level-indexed agent–environment system above induces, at every level i, exactly three distinguishable channels of activity. We define each formally and show that together they exhaust the dynamics of H_i.

Definition 4.4 (Compression Upward) *The* upward channel *at level i is the output process y^{H_i}, which under the coupling relation becomes the input to H_{i+1}:*

$$u^{H_{i+1}}(t) = y^{H_i}(t).$$

Per §1.4, $y^{H_i}(t) \sim g^{H_i}[u^{H_i}, s^{H_i}, t]$, where g^{H_i} maps the full state and input history of H_i to an output. The upward channel is necessarily compressive in the sense already established by the remark on observable parameters from admissible transformations in §1: the parameters of H_i's state that are observable to H_{i+1} are exactly those that survive the quotient induced by admissible transformations of s^{H_i} that leave y^{H_i} unchanged. H_{i+1} does not receive s^{H_i}; it receives only y^{H_i}, a function of s^{H_i} that need not be injective.

Definition 4.5 (Constraint Downward) *The* downward channel *at level i is the input process u^{H_i}, supplied by H_{i+1} under the coupling relation:*

$$u^{H_i}(t) = y^{H_{i+1}}(t).$$

This is the unique channel by which H_{i+1} acts upon H_i: H_{i+1} has no access to H_i's internal state s^{H_i} except as mediated through H_i's own functionals f^{H_i}, g^{H_i} taking u^{H_i} as an argument. The downward channel is therefore exactly as constraining, and no more constraining, than the dependence of f^{H_i} and g^{H_i} on u^{H_i} permits.

Remark 4.6 (Reward as a Special Case of the Downward Channel) *The downward channel above is stated at the level of generality of §4's systems formalism and does not presuppose any particular structure on H_{i+1}'s output. Section 4.2 identifies reinforcement learning as a special case of the optimization model of §2 in which the objective O depends on the history of agent–environment interaction. If H_i is additionally constituted as an optimization model $(\mathcal{H}, O, \mathcal{A})$ per §2, and if O is specified such that $u^{H_i}(t)$ (i.e., $y^{H_{i+1}}(t)$) is interpreted in part as a scalar reward signal entering O, then the downward channel reduces to reward shaping in the ordinary reinforcement learning sense: H_{i+1}'s output constrains H_i not only by entering H_i's state-transition functional f^{H_i} directly, but by entering the objective that H_i's optimization algorithm $\mathcal{A}^{H_i}$ is driven to extremize. This is not a new primitive; it is the downward channel above under the additional structure already available in §2 and §4.2. The general formulation is preferred at the level of the present subsection because it does not require H_i to be constituted as an RL agent specifically; any system per §1 admits a downward channel, whether or not it is also an optimization model.*

Definition 4.7 (Residual Freedom) *Let the state dynamics of H_i be given, in continuous time, by the stochastic differential equation of §1.4:*

$$ds^{H_i}(t) = f^{H_i}_{\text{det}}\big[u^{H_i}, s^{H_i}, t\big]\, dt + \sigma^{H_i}\big[u^{H_i}, s^{H_i}, t\big]\, dW^{H_i}(t),$$

where $f^{H_i}_{\text{det}}$ is the drift term and σ^{H_i} the diffusion coefficient of the (in general u^{H_i}-dependent) stochastic forcing $dW^{H_i}(t)$. This is the explicit decomposition of the general stochastic update $\dot{s}(t) \sim f[u, s, t]$ already given in §1.4, where $\sim$ denotes sampling from a distribution and determinism is recovered as the degenerate case of a distribution with zero variance.

The drift component *$f^{H_i}_{\text{det}}[u^{H_i}, s^{H_i}, t]\,dt$ is the part of H_i's state evolution that is a deterministic function of the downward channel u^{H_i} (and of H_i's own prior state). It is the component of H_i's dynamics that the constraint from H_{i+1} directly compels.*

The diffusion component *$\sigma^{H_i}[u^{H_i}, s^{H_i}, t]\,dW^{H_i}(t)$ is the* residual freedom *of H_i: the part of H_i's state evolution that remains stochastic — irreducible to any deterministic function of u^{H_i} — even after the downward channel is fully specified. Residual freedom is permitted to depend on u^{H_i} through the diffusion coefficient σ^{H_i} itself: the magnitude of H_i's freedom, not only the direction of its compelled drift, may be shaped by the constraint received from above. The discrete-time update rule of §1.4 admits the identical decomposition:*

$$s^{H_i}(t_{j+1}) - s^{H_i}(t_j) = \underbrace{\Delta^{H_i}_{\text{det}}[u^{H_i}, s^{H_i}, t_{j+1}]}_{\textit{compelled}} + \underbrace{\varepsilon^{H_i}[u^{H_i}, s^{H_i}, t_{j+1}]}_{\textit{residual freedom}},$$

where $\Delta^{H_i}_{\text{det}}$ is the deterministic component of f^{H_i} and ε^{H_i} is a zero-mean stochastic term whose distribution may itself depend on u^{H_i}.

Proposition 4.8 (Exhaustiveness of the Three Channels) *For each i, the upward channel (y^{H_i}), the downward channel (u^{H_i}), and residual freedom (the diffusion component of H_i's state update) together account for the complete formal description of H_i as a system per §1: H_i's state process s^{H_i} is exhausted by the drift/diffusion decomposition above, and H_i's interaction with the rest of the stack is exhausted by u^{H_i} and y^{H_i} per the coupling relations of the level-indexed agent–environment system.*

By §1, a system is fully specified by its input process, output process, and (where modeled) state process together with the functionals relating them. The coupling relations of the level-indexed agent–environment system specify u^{H_i} and y^{H_i} in terms of H_i's neighbors in the stack; these are precisely the downward and upward channels. The state process s^{H_i} is specified, per §1.4, by $\dot{s}^{H_i}(t) \sim f^{H_i}[u^{H_i}, s^{H_i}, t]$; the drift/diffusion decomposition of residual freedom is a syntactic rewriting of this same functional into a u^{H_i}-determined component and a residual component, and is exhaustive by construction, since any distribution may be written as a (possibly degenerate) sum of its mean functional and a zero-mean residual. No component of H_i's formal description per §1 falls outside $\{u^{H_i}, y^{H_i}, s^{H_i}\}$, and s^{H_i} is exhausted by drift and diffusion; hence the three channels are exhaustive.

4.3.3 Discussion

The three-channel decomposition above was not imported into the systems formalism of §1–§4; each channel is a direct reading of definitions already present there. The upward channel is the output functional g of §1.3 under the coupling of §4. The downward channel is the input dependence already built into f and g as functionals of u. Residual freedom is the diffusion term already latent in the stochastic differential equation of §1.4, made explicit by a

standard decomposition. What this subsection contributes is not new mathematical content but the recursive application of §4's existing pairwise coupling to a stack of arbitrary depth.

5 Becoming-Held-As-By: Subjects as Systems in Self-Representation

In the language developed in *The Imagination Machine I: A View from Somewhere*, a self-representing subject is an embedded epistemic system whose classifiers appear within its own observation space. The condition that classifiers are themselves observations allows a system to encounter and revise its own acts of classification. The present section approaches the same idea from the perspective of systems modeling: if the formalism developed above can represent any system, then it must also apply to the system performing the representation.

If the above framework above provides insight into how to represent any real system in mathematical terms, then a natural next step is to turn the inquiry on the modeler. In other words, in writing the above formalism I am confronted with the question, "Am I not a real system myself? Can I, then, be understood in these terms?"

I imagine a bubble around my body, and then I imagine it shrinks inward all around and approaches infinitely closely to the edge of my skin. Any measurable passing between this membrane is either input or output–and thus I conceive of agent and environment.

Pursuant to these aims, we now shift from formal system representation to a philosophical and phenomenological inquiry into how a system may represent itself as an agent, co-constituted and co-evolving with an environment. In this way, we move from a formalism for modeling system behavior from an external perspective to a vocabulary by which a self-modeling agentic system may represent its own reality from the internal perspective.

5.1 A Self-Referential Thesis

All may be called the Becoming-Held-As-By[2] (including its becoming held as this by me).

5.2 Potentiality and Representation

Suppose we take "existence" to mean "the quality, state, or event of becoming-held-as-by." We use the word "potentiality" to mean that from which existence emerges through representation. Potentiality is metaphysical substance itself—what we might call the pre-conceptual whatever-I-represent. Representation is the process or result of becoming-held-as-by. A subject is becoming-held-as-by-itself.

For example, to say that a particular cup "exists" is to say that some potentiality (which I could, for example, point to) is becoming held in mind by me as a unified and distinct "thing" which I represent as a cup. If the potentiality does not become held as anything by any subject, then it cannot be said that anything in particular exists there, though there may

[2]The hyphenation of "Becoming-Held-As-By" is deliberate: it reflects the interdependence and co-constitution of the becoming, the holding, the *as*-ness (representation), and the *by*-ness (the subject).

persist some potentiality for becoming-held-as-by (held as a cup, perhaps, or as something else, like a weapon or a hat, by any particular subject). To hold potentiality as something is not to define it once and for all, but to engage in a relationship that may change. The same potentiality may be held as many different things across time, across subjects, or even within the same subject in different moments.

One cannot properly imagine potentiality because all one *can* do is imagine potentiality, in the sense of bringing potentiality into representation. That is to say that to imagine potentiality is already to bring it into representation. Potentiality may have internal structure (e.g. change relative to some internal reference frame according to laws). Regardless, here is the big picture: potentiality (metaphysical substance) is translated into existence (the ontological) through its representation by the subject (the semiological and epistemological: perception, language, systems of meaning, knowledge claims). By this notion of existence, if every conscious being were to disappear suddenly, there would not be a universe at all—only potential for a universe to arise.

Reality, in this account, is enacted through the interplay of potentiality and representation: a process in which potential becomes held through representation, and representation constrains potentiality.

5.3 The Subject Becoming-Held-As-Agent-By-Itself

The most stable world-model I have yet realized is this: world as constituted of agent (self) co-evolving with environment, where the agent's state includes its representation of self, environment, and world; including, recursively, a representation of world as constituted of agent (self) co-evolving with environment, where the agent's state includes its representation of self, environment, and world.

This is a world that I hold as constituted of the agent-environment coupling, wherein a subject may coherently and productively become-held-as-agent-by-itself. The agent is not separable from the environment, though it may be ontologically separate in its own representation. The state of the agent includes its representation of potentiality: It is influenced by its environment's output and its own history, and, in turn, it influences the input to the environment through the output of the agent. Because of the inherent coupling of agent to environment, the subject becoming-held-as-agent-by-itself is to the Becoming-Held-As-By as a *holon* to its greater whole[3]: the subject may become-held-as a distinct object of analysis by itself, and yet it can simultaneously become-held by itself as a part in a larger system.

To use a human-centric analogy, the subject becoming-held-as-agent-by-itself is to the Becoming-Held-As-By as the mouth is to the body: the mouth is not the body, yet it is interconnected with the body; and the declaration that "I am the body" is made by means of the mouth. Similarly, the subject becoming-held-as-agent-by-itself is not the entirety of the Becoming-Held-As-By and yet is embedded (and participating) within it; and the writing and reading of statements like, "All may be called the Becoming-Held-As-By (including its becoming held as this by me)" is enacted by the subject becoming-held-as-agent-by itself.

[3]The term "holon" was coined by Arthur Koestler in his 1967 book, *The Ghost in the Machine*. A holon is both a self-contained entity (hence it is a whole on its own) and at the same time is embedded within a larger containing system or systems (so it is part of a larger whole).

5.4 The Limits of the Systems Formalism

A system is defined by its distinctions: inputs vs. outputs, internal vs. external state. The undivided Whole—that which contains all systems, distinctions, and environments—cannot itself be represented as a system. Since it has no external relation and no boundary, it admits no input/output mapping. Likewise, the complement of the undivided Whole—that is, nothingness, or void—admits no input/output mapping and as such may not be represented as a system.

5.5 Mathematics as Meta-Representation

Mathematics may derive from the structure of representation itself. That is to say, mathematics is a type of meta-representation: a representation of common structure across instances of representation. Accordingly, the representation of mathematical objects could potentially be invariant under change in subject if each subject can in principle abstract from their own instances of representation to arrive at the same mathematical meta-representations. For example, I can map a notion of two-ness to the same symbol that another subject can map theirs, and we can be reasonably sure that we agree on its meaning, because we both experience unity and difference in our representations of self and world. Likewise, I can map a notion of a function to the same symbol that another can map theirs, and we can be reasonably sure that we agree on its meaning, because we both represent and abstract from instances of change. Unity, difference, and change may be necessary structures of subjective representation, such that any subject with sufficient abstract reasoning capability can attribute the same meaning to the same meta-representations.

5.6 Haecceity and Qualia

Complementary to the notion of meta-representation in this account is the notion of haecceity, or the irreducible *this*-ness of an entity. Haecceity is what remains in representation modulo meta-representation—the particularity that is not captured by abstraction from representation to meta-representation. For a human, haecceity may correspond to the irreducible qualia of the experience of being *this particular self* in *this particular moment*.

5.7 Truth and Coherence

A proposition is a linguistic claim that may be judged true or false by a subject. Truth is a judgment of coherence among a collection of propositions. Formally, a proposition is judged false if it is shown that the proposition, potentially together with a collection of propositions judged to be true, implies contradiction of a proposition judged true. Therefore, a particular proposition is judged true only by virtue of its ongoing coherence with a collection of mutually non-contradictory propositions. It must be emphasized that propositions involving instantiation (of abstract classes) are among the propositions that must be coherent in a collection of truths. For example, a proposition like, "An electron evolves according to the Schrodinger equation," must cohere with such propositions of instantiation as, "This reading (referring to a particular representation in experience) is due to an electron," and, "This

reading (at another time, perhaps) is not due to an electron," as well as propositions that are not instantiations like, "An electron has negative charge." This understanding of truth allows for pluralism while requiring that a worldview be consistently tethered to moments of becoming-held-as-by.

5.8 Conclusion

The central claim is that we are always describing the world from the inside: embedded within the Becoming-Held-As-By and co-evolving with our environment, seeing patterns in our seeing-patterns. We conscious beings are individually and collectively a self-representing network of interacting holonic subsystems. And yet, on the whole and within each part, haecceity remains.

Linear Encoding of Koopman Spectra in Predictive Agents

Mark Tracy
Boston University
mrktracy@bu.edu

Abstract

We study whether a predictive agent trained on prediction error alone encodes the Koopman spectral structure of its environment in a linearly decodable form in its weights. The answer is yes, and it holds across qualitatively distinct environment classes.

The agent architecture is fixed throughout: a recurrent MLP with a bottleneck layer (hidden dimension $d_h = d_s = 32$, compressing the concatenated state–observation input of dimension $d_s + d_o = 38$), observation layer normalization, and a linear prediction head. All agents share a common weight initialization derived from a fixed random seed; any difference in converged parameters is attributable solely to the environment.

In the first setting, the agent is embedded in a quasi-periodic environment consisting of three oscillators with incommensurate frequencies, observed only through relational phase differences. Frequency vectors are sampled from the uniform Dirichlet$(1, 1, 1)$ distribution over the normalized simplex — the maximum-entropy prior over all quasi-periodic environments of this form — grounding a universality claim. A randomly drawn same-size slice of the converged parameter vector linearly encodes the complete independent Koopman spectrum with joint $R^2 = 0.968 \pm 0.003$ under a strictly linear probe across $N = 5{,}000$ independent agents — a result that holds uniformly across $P = 20$ independent random draws of the slice, establishing that the encoding is distributed across the full parameter vector rather than localized to any specific layer.

Via a canonical cyclic lifting, the same architecture and the same mechanism extend to arbitrary sequences over finite dictionaries. For deterministic periodic sequences (unit fractions $1/n$ with known periods), the architecture recovers three Koopman modes with joint $R^2 = 0.999 \pm 0.000$; a paired null experiment on the decimal expansion of π returns joint $R^2 = -0.002 \pm 0.000$, confirming that the result depends on spectral structure in the sequence and not on probe design. For a nearest-neighbor random walk on $\mathbb{Z}/10\mathbb{Z}$ parameterized by a continuously varying asymmetry q, the architecture recovers the Koopman eigenvalues with joint $R^2 = 0.981 \pm 0.001$, with a smooth continuous readout rather than discrete clustering.

The inter-instance probe is geometrically consistent within a shared initialization: a single linear map trained on the agent population predicts each environment's Koopman spectrum from the corresponding agent's parameter slice. This consistency is relative to the fixed initialization coordinate frame; agents sharing a common initialization converge to parameter vectors whose inter-environment variation tracks the Koopman spectrum along consistent linear directions throughout.

Contents

1 Naturalistic Motivation

A long-standing question in representation learning is whether agents that predict well necessarily represent their environments in a structured, interpretable way. The Koopman operator framework [1, 2] offers a precise version of this question: an environment with periodic or quasi-periodic dynamics has a Koopman spectrum — a set of eigenvalues characterizing its rotation rates — that is, in principle, the right thing for a predictive agent to encode. Does a predictive agent, trained only on prediction error, actually encode this spectrum? And if so, does it encode it in a linearly accessible form?

Recent work on data-driven Koopman methods [3, 4, 5] has pursued learned Koopman embeddings through explicit operator and decoder losses: the network is trained to find coordinates in which dynamics are linear, with the linearity enforced directly as an objective. Our setting differs in two ways. First, the agent is trained only on next-step prediction error; there is no explicit Koopman loss or decoder. Second, we probe the converged parameter vector directly rather than learning an embedding. This lets us ask a different question: not "can we build a network that linearizes the dynamics," but "does a network trained to predict incidentally organize its full parameter vector so that the dynamics are linearly readable?"

The probing methodology we use is related to the diagnostic probing literature in NLP [6, 7], where linear probes on network representations test whether specific information is linearly encoded. Our setting differs in an important respect: we probe the *weights* of the converged agent rather than the activations of a trained network on a dataset. The probe is applied to randomly drawn same-size slices of the full parameter vector; the uniformity of the result across independent random draws establishes that the Koopman encoding is a global property of the solution, not an artifact of probing any particular location.

Crucially, the probe is an *inter-instance* test. After training N independent agents, each on a distinct environment, we train a single linear map that predicts the Koopman eigenvalue of each environment from the corresponding agent's parameter slice. High R^2 means that the encoding is *geometrically consistent across agents*: the same linear map works for all of them simultaneously. This consistency is established relative to a fixed initialization: all agents share a common weight initialization derived from a fixed random seed, so any difference in converged parameters is attributable solely to the environment. This is a structural claim about the solution geometry that SGD converges to, not a claim about any individual agent's internal representation.

The paper is organized as follows. Section 2 introduces the quasi-periodic environment and its Koopman structure. Section 3 describes the predictive agent architecture and its theoretical motivation. Section 4 develops the Koopman connection formally. Section 5 presents the empirical protocol, including the inter-instance probing interpretation. Section 6 reports results for the quasi-periodic experiment. Section 7 introduces the canonical cyclic lifting that extends the architecture to arbitrary sequences over finite dictionaries. Sections 8–10 develop the Koopman structure of periodic sequences, the null experiment, and stochastic environments. Section 11 presents the extended experimental results. Section 12 discusses theoretical implications and future directions.

2 The Quasi-Periodic Environment

The quasi-periodic environment provides the cleanest initial setting: a system with a known, analytically tractable Koopman spectrum and a naturalistic motivation.

Environments of this type arise whenever an embedded observer tracks multiple periodic processes with incommensurate frequencies. The Earth–Sun–Moon system is the canonical example:

the solar day, lunar cycle, and solar year have periods whose ratios are not rational, so their relative phases drift continuously and never exactly repeat. An observer embedded in such a system confronts incommensurate cycles whose relative phases must be estimated from observation rather than computed from exact knowledge of the underlying frequencies. There arises infinite novelty in relational observables from ordered generative structure.

The Earth-Sun-Moon system is quasi-periodic. There isn't a perfectly rational number of days in a month, nor months in a year. The consequence is that inductive estimation becomes useful for survival on Earth. To see why, consider the two edge cases: if the universe as experienced by an Earth-bound observer were perfectly random, there would be no selective pressure for any sort of cognition — prediction would be useless by definition. Conversely, if the configuration of the universe eventually repeated perfectly, a body need only synchronize mechanically to the eternally repeating rhythm to achieve stability. The quasi-periodic system sits exactly between these extremes, rewarding inductive estimation of the underlying geometric structure of the observed relations within the universe.

Let the environment consist of three cyclic variables

$$\theta_1(t),\, \theta_2(t),\, \theta_3(t) \in S^1 \cong \mathbb{R}/2\pi\mathbb{Z},$$

with dynamics

$$\theta_i(t+1) = \theta_i(t) + \omega_i \pmod{2\pi}, \qquad i = 1, 2, 3.$$

Definition 1 (Quasi-Periodic System)**.** *The system is quasi-periodic if $\omega_1, \omega_2, \omega_3$ are rationally independent, i.e., the only integer solution to $k_1\omega_1 + k_2\omega_2 + k_3\omega_3 = 0$ is $k_1 = k_2 = k_3 = 0$. In this case the trajectory is dense on the torus $\mathbb{T}^3 = S^1 \times S^1 \times S^1$.*

2.1 Observation Model

The agent observes only relational quantities $o_t = h(x_t)$, where $x_t = \theta(t)$ and

$$h(x_t) = (\cos\Delta_{12},\, \sin\Delta_{12},\, \cos\Delta_{13},\, \sin\Delta_{13},\, \cos\Delta_{23},\, \sin\Delta_{23}),$$

with $\Delta_{ij}(t) = \theta_i(t) - \theta_j(t) \pmod{2\pi}$. The six-dimensional observation vector is the real and imaginary decomposition of the three complex Koopman eigenfunctions $z_{ij}(t) = e^{i\Delta_{ij}(t)}$.

By computing relational phase differences and lifting to the unit circle, the observation function places the dynamics in a space where they are linear — the same coordinate choice pursued explicitly in data-driven Koopman methods [3, 4]. Given these observables, prediction error alone drives the Koopman spectral structure into a linearly decodable form in the agent's parameters.

2.2 Environment Distribution

Frequency vectors are sampled from the interior of the normalized simplex $\Delta^2 = \{\omega \in \mathbb{R}^3 : \omega_i > 0,\ \omega_1 + \omega_2 + \omega_3 = 1\}$ via the uniform Dirichlet$(1, 1, 1)$ distribution.

This is the maximum-entropy distribution over the simplex: it places equal probability on all frequency compositions, imposing no prior preference for any particular regime. Sampling from Dirichlet$(1, 1, 1)$ is equivalent to sampling uniformly over all quasi-periodic environments of this form, grounding the universality claim: the result holds not for a hand-tuned distribution but across the full space of environments. Rational independence is guaranteed almost surely under continuous sampling.

3 The Predictive Agent

A predictive agent is defined by three functions:

$$o_t = h(x_t), \qquad s_{t+1} = u(s_t, o_t), \qquad \hat{o}_{t+1} = g(s_{t+1}, o_t),$$

where s_t is internal state and $\hat{o}_{t+1}$ is the predicted next observation.

3.1 Neural Parameterization

The observation is normalized via a local layer normalization:

$$\mathring{o}_t = \mathrm{LayerNorm}(o_t).$$

This operation is strictly local and causal: it normalizes the six observation channels at each timestep independently, using only o_t, with no access to past or future. It removes within-timestep scale variation while preserving relative magnitudes of the phase difference components.

The state update is parameterized by a residual connection with layer normalization:

$$s_{t+1} = \mathrm{LayerNorm}(s_t + \mathrm{MLP}_u([s_t,\, \mathring{o}_t])) \,.$$

The MLP computes a compact update $\Delta s_t = \mathrm{MLP}_u([s_t, \mathring{o}_t])$, added to the current state before normalization. This residual form implements a delta update: the network learns what needs to change at each timestep rather than reconstructing the full state from scratch. Layer normalization after the residual addition prevents unbounded accumulation over the trajectory.

The MLP consists of two hidden layers of dimension d_h, decomposing the state update into three linear operations:

$$\mathrm{Linear}(d_s + d_o \to d_h) \to \mathrm{ReLU} \to \mathrm{Linear}(d_h \to d_h) \to \mathrm{ReLU} \to \mathrm{Linear}(d_h \to d_s).$$

The first map compresses the concatenated state–observation input of dimension $d_s + d_o = 38$ into a perceptual representation of dimension $d_h = 32 < 38$; the third expands back into a state-dimension update. The compression from $d_s + d_o = 38$ to $d_h = 32$ may serve as a form of regularization, discouraging high-dimensional nonlinear solutions and encouraging the persistent environmental structure to be encoded in a compact, linearly accessible form throughout the parameter vector.

The prediction head is a linear readout:

$$\hat{o}_{t+1} = W[\, s_{t+1},\, \mathring{o}_t \,] + b.$$

This decomposition is structurally significant. The normalized observation $\mathring{o}_t$ provides the current phase position $(\cos\Delta_{ij}(t), \sin\Delta_{ij}(t))$ directly; the state s_{t+1} must supply the environment-specific rotation amounts $e^{i(\omega_i - \omega_j)}$, which are constant across trajectories. Since the prediction head is linear, it extracts these rotation amounts as linear projections of the state. The state must therefore encode the rotation amounts in a linearly accessible form.

3.2 Independent Agent Training

Each environment k is assigned an independent agent initialized from identical weight vectors ϕ_0 derived from a fixed random seed. Each agent trains exclusively on its assigned environment for N_{epochs} epochs, each consisting of a fresh trajectory of length T with a randomly drawn initial phase. The use of fresh trajectories at each epoch is significant: trajectory-specific content cannot be reliably encoded in the weights, since the sequence is never the same twice. Any difference in the converged parameter vector $\phi^{(k)}$ across agents is attributable to the persistent dynamical structure of environment k — the rotation rates, period, or asymmetry — rather than to any particular sequence of observations.

3.3 Prediction Error and Training

Training minimizes

$$\mathcal{L} = \frac{1}{T}\sum_{t=0}^{T-1} \|\hat{o}_{t+1} - o_{t+1}\|_2^2,$$

averaged over trajectory timesteps. Parameter updates use Adam with gradient clipping at unit global norm.

4 Observable Invariants and the Koopman Connection

4.1 Time-Rescaling Symmetry

Proposition 1. *The frequency vector ω is identifiable only up to multiplication by a positive scalar from relational phase observations alone.*

Proof. For $\lambda > 0$, define $\omega' = \lambda\omega$. The dynamics of the rescaled system are

$$\theta'_i(t+1) = \theta'_i(t) + \lambda\omega_i \pmod{2\pi}.$$

After t steps from initial condition $\theta'(0) = \theta(0)$,

$$\theta'_i(t) = \theta_i(0) + \lambda\omega_i t \pmod{2\pi}.$$

The pairwise differences observed by the agent are

$$\Delta'_{ij}(t) = \theta'_i(t) - \theta'_j(t) = (\omega_i - \omega_j)\lambda t \pmod{2\pi}.$$

The original system gives

$$\Delta_{ij}(t) = (\omega_i - \omega_j)t \pmod{2\pi}.$$

Since the observation function h depends only on $(\cos\Delta_{ij}(t), \sin\Delta_{ij}(t))$, two trajectories produce identical observations at all t if and only if $\Delta'_{ij}(t) \equiv \Delta_{ij}(t') \pmod{2\pi}$ for some time correspondence $t \leftrightarrow t'$. Taking $t' = \lambda t$ gives $\Delta'_{ij}(t) = \Delta_{ij}(\lambda t)$, so the trajectory of ω' is a time-reparameterization of the trajectory of ω and no relational observation distinguishes them. □

Observable invariants are therefore the projective equivalence class $[\omega_1 : \omega_2 : \omega_3]$. Normalizing via $\omega_1 + \omega_2 + \omega_3 = 1$, all relationally distinguishable environments correspond to points in the interior of Δ^2.

4.2 Koopman Representation

Writing $z_{ij}(t) = e^{i\Delta_{ij}(t)}$, the relational dynamics imply

$$z_{ij}(t+1) = e^{i(\omega_i - \omega_j)}\, z_{ij}(t).$$

The observable evolves by multiplication by a constant complex phase factor: a Koopman eigenfunction. The nonlinear state dynamics on $\mathbb{T}^3$ become linear in the space of relational observables.

For each observable pair (i, j), accurate prediction requires

$$\cos(\Delta_{ij}(t+1)) = \cos(\Delta_{ij}(t))\cos(d_{ij}) - \sin(\Delta_{ij}(t))\sin(d_{ij}),$$

where $d_{ij} = \omega_i - \omega_j$. The phase position is available from o_t; the rotation amounts $\cos(d_{ij})$ and $\sin(d_{ij})$ must come from s_{t+1}. The prediction head extracts these as linear projections of the state; they must therefore be linearly encoded somewhere in the agent's parameters.

4.3 Independence of the Koopman Spectrum

Proposition 2 (Independence of the Koopman Spectrum)**.** *The three pairwise Koopman eigenvalues $e^{id_{12}}, e^{id_{13}}, e^{id_{23}}$ satisfy $e^{id_{13}} = e^{id_{12}} \cdot e^{id_{23}}$, so the spectrum has two degrees of freedom. Sorting $\omega_{\min} \leq \omega_{\text{mid}} \leq \omega_{\max}$, define $d_{\text{span}} = \omega_{\max} - \omega_{\min}$, $d_{\text{least}} = \min(\omega_{\text{mid}} - \omega_{\min}, \omega_{\max} - \omega_{\text{mid}})$, and $d_{\text{mid}} = \max(\omega_{\text{mid}} - \omega_{\min}, \omega_{\max} - \omega_{\text{mid}})$. Then $(d_{\text{mid}}, d_{\text{least}})$ is a complete independent basis for the Koopman spectrum.*

Proof. The composition law follows from $d_{13} = d_{12} + d_{23}$, which implies $e^{id_{13}} = e^{id_{12}} \cdot e^{id_{23}}$. The greatest pairwise difference $d_{\text{span}} = d_{\text{least}} + d_{\text{mid}}$ carries no independent information. □

Remark 1. *The composition law reflects the group structure of $U(1)$: the three pairwise eigenvalues form a generating set with one relation. Probing all three introduces redundancy; the canonical probe targets the two generators $(d_{\text{mid}}, d_{\text{least}})$.*

5 Empirical Protocol

5.1 Linear Probing as an Inter-Instance Structural Test

After training on environment k, the full converged parameter vector $\phi^{(k)} \in \mathbb{R}^D$ is retained, where $D = 3{,}670$ is the total number of agent parameters. For each of $P = 20$ independent trials, a random slice $\psi^{(k)} \in \mathbb{R}^d$ of fixed size $d = 734$ (20% of the total parameters) is drawn uniformly without replacement from the full parameter vector. Slices are in general non-contiguous. We train a linear probe

$$\hat{y} = W\psi + b$$

to predict the Koopman eigenvalue components from the parameter slice, using mean squared error with ℓ_2 regularization.

The central empirical finding is that this probe succeeds uniformly across independent random draws of the slice. This establishes that the Koopman encoding is a global property of the converged parameter vector: the inter-environment variation in $\phi^{(k)}$ tracks the Koopman spectrum of environment k along consistent linear directions throughout the parameter space, not at any privileged location within it. We report the mean and standard deviation of Joint R^2 across the $P = 20$ draws; near-zero variance across draws is the primary evidence for distributed encoding.

It is important to be precise about what this probe tests. Each agent has been trained on a single environment with a single distinguishing constant (the rotation rates, or later the period or asymmetry parameter). The claim is that the encoding is *geometrically consistent across agents*: a single linear map, trained on the agent population, predicts the eigenvalue from the parameter slice across all environments simultaneously. This is a structural claim about the solution geometry SGD converges to, relative to a fixed initialization coordinate frame. Even if each agent has perfectly encoded its environment's constant, those encodings could in principle be organized arbitrarily — idiosyncratic rotations, scalings, or nonlinear solutions that differ across agents. High R^2 under a single linear map means they are not: the solution space organizes itself so that inter-agent variation in the parameters tracks inter-environment variation in the Koopman spectrum along consistent linear directions.

The use of a strictly linear probe is a structural commitment, not a simplification. A nonlinear probe could recover eigenvalue information from any representation; it would tell us nothing about organization. The linear probe tests specifically whether the Koopman eigenvalue components have been encoded linearly and consistently across the agent population.

5.2 Probe Targets

For the quasi-periodic experiment, the probe targets are the real and imaginary parts of two independent Koopman eigenvalues:

$$y^{(k)} = \big(\cos(d^{(k)}_{\text{mid}}),\ \sin(d^{(k)}_{\text{mid}}),\ \cos(d^{(k)}_{\text{least}}),\ \sin(d^{(k)}_{\text{least}})\big).$$

The largest interval d_{span} is excluded as it is algebraically determined by d_{mid} and d_{least}.

5.3 Generalization and Robustness

Generalization is evaluated via 5-fold cross-validation over the agent pool. The weight vector normalizer (StandardScaler) is refit on the training split within each fold to prevent leakage. Probe performance is reported as mean ± std R^2 across folds, averaged across the $P = 20$ random slice draws.

Strong mean R^2 with small fold variance indicates that the parameter vector captures general dynamical invariants of the environment class. The R^2-vs-α curve provides a further diagnostic: a broad plateau indicates a stable invariant accessible to a wide range of linear readouts, rather than a fragile signal requiring precise tuning. The *shape* of this curve is as informative as the R^2 value itself, and we report it for all experimental conditions, overlaying the individual curves for each of the $P = 20$ random draws. Near-zero spread across those curves is the primary visual evidence for distributed encoding.

6 Results: Quasi-Periodic Experiment

We report results from $N = 5{,}000$ independent agents trained on environments sampled from Dirichlet$(1, 1, 1)$, with trajectory length $T = 2{,}000$, $N_{\text{epochs}} = 200$, and the canonical architecture ($d_h = d_s = 32$, observation LayerNorm). The trajectory length $T = 2{,}000$ ensures identifiability: environments with a slow oscillator require long trajectories for the rotation rate to be reliably encoded. Generalization is evaluated via 5-fold cross-validation across $P = 20$ random parameter slices of size $d = 734$.

Quantity	**Mean**	**Std**
Joint R^2 across 20 slices	0.968	0.003
Min slice Joint R^2	0.963	—
Max slice Joint R^2	0.974	—

The mean Joint $R^2 = 0.968 \pm 0.003$ across $P = 20$ random parameter slices establishes that the complete independent Koopman spectrum is linearly encoded in the converged parameter vector across the full Dirichlet$(1, 1, 1)$ distribution — the uniform prior over all quasi-periodic environments of this form. The range across slices (0.963–0.974) and near-zero standard deviation confirm that the encoding is not concentrated at any privileged location: any 20% window into the parameter vector recovers the spectrum with essentially the same fidelity. The R^2-vs-α curves for the 20 draws are visually indistinguishable throughout the plateau, providing a second confirmation of uniformity. The distributed encoding is discussed further in Section 12.

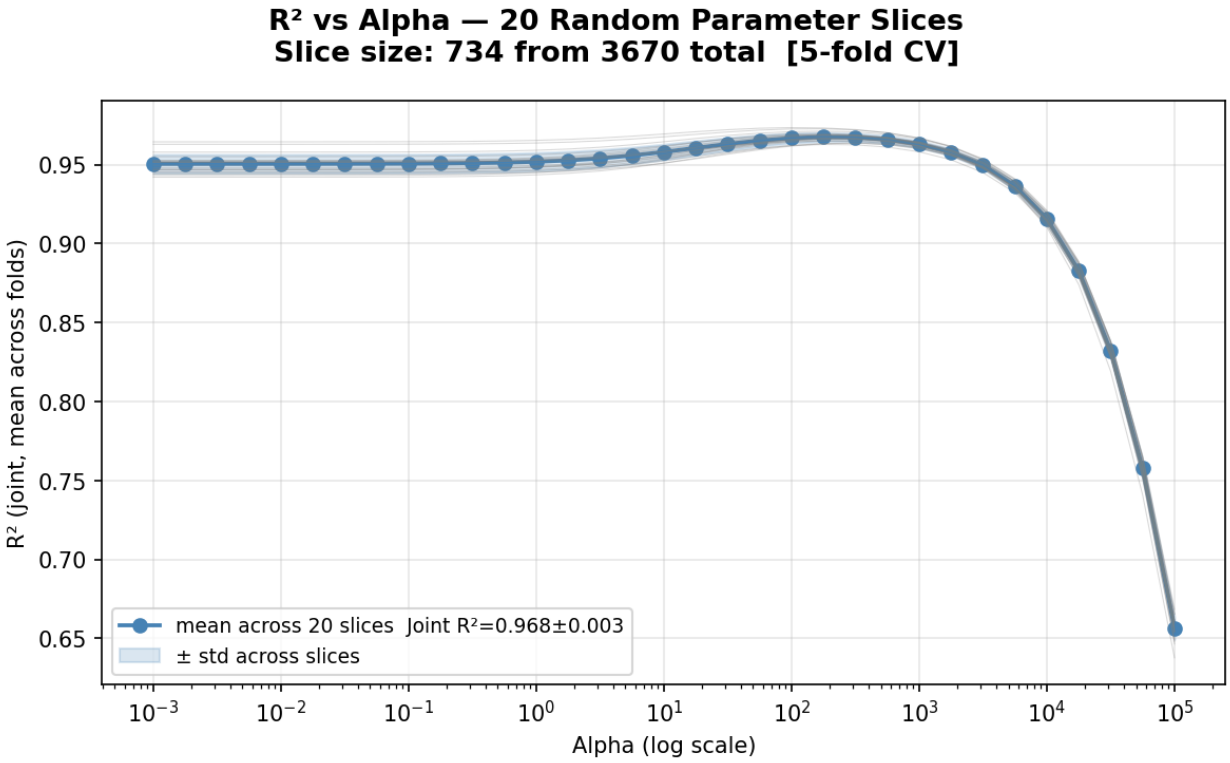

Figure 1: R^2 **vs.** α **curves, quasi-periodic experiment.** Mean Joint $R^2 \pm$ std across 5 folds as a function of regularization strength α, for each of $P = 20$ independently drawn random parameter slices (grey curves) and their mean (blue). The curves are nearly indistinguishable throughout the plateau region, confirming that linear decodability of the Koopman spectrum is uniformly distributed across the parameter vector. Peak mean Joint $R^2 = 0.968 \pm 0.003$. $N = 5{,}000$, $T = 2{,}000$, $d = 734$ from $D = 3{,}670$.

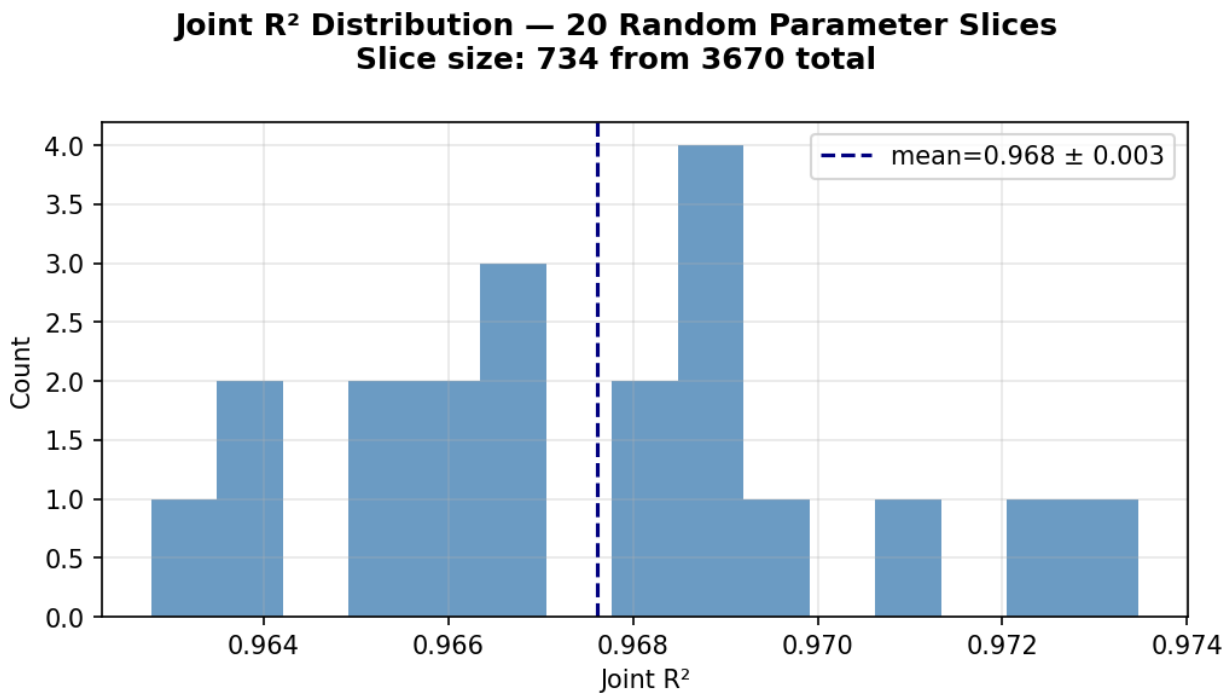

Figure 2: **Distribution of Joint R^2 across 20 random parameter slices, quasi-periodic experiment.** Each bar corresponds to one random draw of $d = 734$ parameters from the full $D = 3{,}670$-dimensional parameter vector. The distribution is tightly concentrated (range 0.963–0.974, std = 0.003), confirming that Koopman spectral structure is recoverable from any same-size window into the parameter vector.

7 The Canonical Cyclic Lifting

The quasi-periodic experiment establishes the existence result in the cleanest possible setting. We now ask how general it is. The key observation is that any data stream over a finite dictionary of size M is indexed by $\mathbb{Z}/M\mathbb{Z}$. This motivates a canonical lifting procedure that converts any such sequence into an observation vector of exactly the form the architecture processes, with no modification required beyond a fixed phase offset drawn once before training to ensure non-degeneracy. The lifting takes the finite alphabet as a cyclic index set and embeds pairwise differences into the unit circle; whether spectral structure is present in the lifted representation depends on the data, and is determined empirically rather than assumed. The π null experiment demonstrates this directly: the lifting is identical for π and unit fractions, but no spectral structure is recoverable from π.

7.1 Cyclic Groups and the Character Map

Definition 2 (Cyclic Group Sequence)**.** *A cyclic group sequence of order M is a sequence $(z_t)_{t\geq 0}$ with $z_t \in \mathbb{Z}/M\mathbb{Z}$ for all t.*

The canonical character $\chi : z \mapsto e^{2\pi i z/M}$ is the unique faithful group homomorphism of minimal order from $\mathbb{Z}/M\mathbb{Z}$ to $U(1)$. It satisfies $\chi(z+y) = \chi(z)\chi(y)$ for all z, y, where addition is modulo M.

7.2 Sequences Over Finite Dictionaries

Consider a sequence $(a_t)_{t\geq 0}$ over a finite dictionary, A, i.e. $a_t \in A$ for all t, where A is a set with finite cardinality M. By establishing an arbitrary ordering of members of A, we define a dictionary, an ordered tuple $P = (p_1, \ldots, p_M)$, where $\forall p_i \in P \quad \exists t \geq 0 : p_i = a_t$. We then define an embedding of the dictionary into a cyclic group by means of an index mapping: $f : P \to \mathbb{Z}/M\mathbb{Z} : p_i \mapsto i$. This allows us to define a sequence $(x_t)_{t\geq 0}$, where $x_t = f(p_i)$, where $p_i = a_t$.

7.3 Pairwise Differences and the Observation Vector

Given three consecutive sequence values (x_t, x_{t+1}, x_{t+2}), compute pairwise modular differences in $\mathbb{Z}/M\mathbb{Z}$:

$$\begin{aligned} d_{12}(t) &= (x_t - x_{t+1}) \bmod M, \\ d_{13}(t) &= (x_t - x_{t+2}) \bmod M, \\ d_{23}(t) &= (x_{t+1} - x_{t+2}) \bmod M, \end{aligned}$$

with composition law $d_{13}(t) = (d_{12}(t) + d_{23}(t)) \bmod M$.

Lift each difference to the unit circle: $\theta(d_{ij}) = 2\pi d_{ij}/M \pmod{2\pi}$.

Definition 3 (Canonical Cyclic Lifting)**.** *Let $\varphi_0 \sim \text{Uniform}(0, \pi/4)$ be a fixed phase offset drawn once before training. The canonical cyclic lifting produces:*

$$o_t = \big(\cos\theta(d_{12}),\, \sin\theta(d_{12}),\, \cos\theta(d_{13}),\, \sin\theta(d_{13}),\, \cos\theta(d_{23}),\, \sin\theta(d_{23})\big) \in \mathbb{R}^6,$$

where $\theta(d_{ij}) = 2\pi d_{ij}/M + \varphi_0 \pmod{2\pi}$. The phase offset ensures that both cosine and sine components are informative for all M, including powers of 2 where the unshifted lifting is degenerate. This is analogous in structure to the quasi-periodic observation vector of Section 2. The architecture applies without modification.

The lifting requires knowing M but no other structural knowledge of the sequence. For any discrete data stream over a known finite dictionary, M is the dictionary size. Examples include $M = 10$ for decimal digit sequences, $M = 256$ for byte streams, and $M = 12$ for chromatic pitch sequences. The lifting is canonical given M and requires no design choices beyond it.

7.4 Universality of the Cyclic Observation Space

The trajectory of a 3-body quasi-periodic system is dense on the torus. This implies that the differences between relative phases of consecutive oscillator pairs, taken modulo 2π, never ultimately terminate or fall into repetition. Therefore any finite sequence over a cyclic group can be approximated arbitrarily closely by this observable of the relational dynamics at some point in the trajectory. By lifting any sequence indexed by a cyclic group to the unit circle, we exploit this mapping — the cyclic observation space is a universal coordinate system for finite sequences, and the agent's converged parameters extract the spectral fingerprint of the corresponding quasi-periodic trajectory, wherever such structure exists in the lifted representation.

8 Koopman Structure of Periodic Sequences

Proposition 3 (Koopman Eigenvalues of a Period-p Sequence)**.** *Let (x_t) be a periodic sequence over $\mathbb{Z}/M\mathbb{Z}$ with period p. The Koopman eigenvalues of the canonically lifted sequence are the p-th roots of unity $\lambda_k = e^{2\pi ik/p}$, $k = 0, \ldots, p-1$, with independent modes $k = 1, \ldots, \lfloor p/2 \rfloor$.*

Proof. A periodic sequence with period p defines a cyclic dynamical system on a p-point orbit. The Koopman operator satisfies $\mathcal{K}^p = \mathrm{Id}$, so its eigenvalues satisfy $\lambda^p = 1$, giving the p-th roots of unity. □

8.1 Unit Fractions as Environments

For $1/n$ with $\gcd(n, 10) = 1$, the period of the decimal expansion is $p = \mathrm{ord}_n(10)$, the multiplicative order of 10 modulo n. Different values of n give different periods and therefore different Koopman spectra. The period p plays exactly the role that the rotation rate ω plays in the quasi-periodic experiment: it is the single environment-distinguishing variable encoded in the agent's parameters from prediction error alone. The probe targets are computable before any training begins.

n	Period p	Fundamental eigenvalue $e^{2\pi i/p}$
7	6	$e^{i\pi/3}$
11	2	$e^{i\pi}$
13	6	$e^{i\pi/3}$
17	16	$e^{i\pi/8}$
19	18	$e^{i\pi/9}$
23	22	$e^{i\pi/11}$
29	28	$e^{i\pi/14}$
31	15	$e^{2i\pi/15}$

9 The Null Experiment

Definition 4 (Normal Sequence)**.** *A sequence (x_t) over $\mathbb{Z}/M\mathbb{Z}$ is normal if every block of length n occurs with limiting frequency M^{-n}, so the one-step transition distribution is uniform: $p(y \mid x) =$*

$1/M$ *for all* x, y.

Proposition 4 (Normality Implies No Recoverable Spectral Structure)**.** *If* (x_t) *is normal, the Fourier spectrum of the lifted sequence is flat and the Koopman spectrum is trivial. A linear probe on the parameter vector from agents trained on this sequence should return* $R^2 \approx 0$.

The null experiment uses a **paired design**. The agent pool is constructed with the same denominators n, the same periods p, and therefore the same probe targets as the unit-fraction pool. Only the digit sequence differs: each agent trains on a non-overlapping window of the decimal digits of π rather than the periodic expansion of $1/n$. Any difference in probe R^2 between the two experiments is attributable entirely to the presence or absence of Koopman structure in the digit sequence, not to any difference in probe design, target distribution, or architecture.

Conjecture 1 (Normality of π, [8])**.** *The decimal expansion of* π *in base 10 is a normal sequence over* $\mathbb{Z}/10\mathbb{Z}$.

This conjecture is widely believed but unproven. We use π as a null control because it is the canonical candidate for a structureless digit sequence over $\mathbb{Z}/10\mathbb{Z}$. The near-zero R^2 returned by a probe specifically calibrated to detect spectral structure constitutes weak empirical evidence for normality in the windows tested; normality is a limiting property and no finite experiment can establish it.

10 Stochastic Environments: The Nearest-Neighbor Random Walk

10.1 The Random Walk Model

Definition 5 (Nearest-Neighbor Random Walk)**.** *A nearest-neighbor random walk on* $\mathbb{Z}/M\mathbb{Z}$ *with asymmetry parameter* $q \in (0, 1)$ *is the Markov chain with* $P_{i,(i+1) \bmod M} = q$, $P_{i,(i-1) \bmod M} = 1 - q$, *and* $P_{ij} = 0$ *otherwise.*

This is a circulant chain with uniform stationary distribution. The parameter q governs the asymmetry: $q = 1/2$ gives a symmetric walk; $q \neq 1/2$ introduces drift.

10.2 Koopman Eigenvalues of the Random Walk

Proposition 5. *The eigenvalues of the transition matrix are*

$$\lambda_k = \cos(2\pi k/M) + i(2q - 1)\sin(2\pi k/M), \qquad k = 0, \ldots, M - 1.$$

Proof. P is circulant with first row $(0, q, 0, \ldots, 0, 1 - q)$. Its eigenvalues are the discrete Fourier transform of the first row. Applying Euler's formula gives the result. □

Remark 2. *The real parts* $\cos(2\pi k/M)$ *are constant across environments and carry no information about* q. *The imaginary parts* $(2q-1)\sin(2\pi k/M)$ *vary continuously and monotonically with* q *and constitute the probe targets. The parameter* q *plays the role of the period* p *in the unit-fraction experiment: it is the single environment-distinguishing variable. In contrast to the unit-fraction experiment,* q *varies continuously, so the probe operates as a genuine regression rather than approximate classification.*

11 Experimental Protocol and Results: Cyclic Lifting

11.1 Architecture and Data Generation

The predictive agent is identical to the quasi-periodic experiment. No modification to the architecture is required for any of the following experimental settings.

Unit fractions. N agents, each assigned $1/n$ drawn cyclically from $\{7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47\}$. Decimal expansions tiled to length $T + 3$; window offsets cycle through $\{0, \ldots, p - 1\}$.

π null. Same denominators and periods, giving identical probe targets. Non-overlapping windows of length $T + 3$ from the decimal expansion of π replace the fraction sequences.

Nearest-neighbor random walk. Each environment assigned $q_i \sim \text{Uniform}(0.01, 0.99)$. Trajectories of length $T + 3$ on $\mathbb{Z}/10\mathbb{Z}$ from uniformly drawn initial states.

11.2 Probe Targets

Unit fractions and π. $y^{(k)} = (\cos(2\pi/p_k), \sin(2\pi/p_k), \ldots) \in \mathbb{R}^{2N_{\text{modes}}}$, computed in closed form before training begins.

Nearest-neighbor random walk. $y^{(i)} = ((2q_i{-}1)\sin(2\pi/M), (2q_i{-}1)\sin(4\pi/M), \ldots) \in \mathbb{R}^{N_{\text{modes}}}$.

Probe targets are StandardScaled per channel within each fold prior to Ridge fitting and inverse-transformed for R^2 computation.

11.3 Results: Unit Fractions

$N = 5{,}000$ agents, $T = 200$, $N_{\text{epochs}} = 100$, $N_{\text{modes}} = 3$, single initialization, $P = 20$ random slices.

Quantity	Mean	Std
Joint R^2 across 20 slices	0.999	0.000
Min slice Joint R^2	0.999	—
Max slice Joint R^2	0.999	—

All 20 random slices achieve Joint $R^2 = 0.999 \pm 0.000$, with a range of only 0.9986–0.9992. The R^2-vs-α curves for all 20 draws are visually indistinguishable, forming a single flat line at ceiling across a six-decade plateau. The near-perfect recovery is partly a reflection of the discrete structure of the environment pool; the continuous-parameter random walk experiment below provides the more demanding test of genuine regression.

11.4 Results: π Null Experiment

Same structure as unit fractions; π digits replace fraction sequences.

Joint $R^2 = -0.002 \pm 0.000$ across all 20 random slices; every draw returns a value in the range -0.0026 to -0.0018. The R^2 curve never plateaus for any slice — it rises monotonically toward zero as $\alpha \to \infty$, with best $\alpha = 100{,}000$ uniformly. Every π agent converged to a structureless solution (symmetric loss distribution at mean = 0.285): the mean predictor of a normal sequence. The paired design makes the comparison exact: the difference 0.999 versus -0.002 is attributable

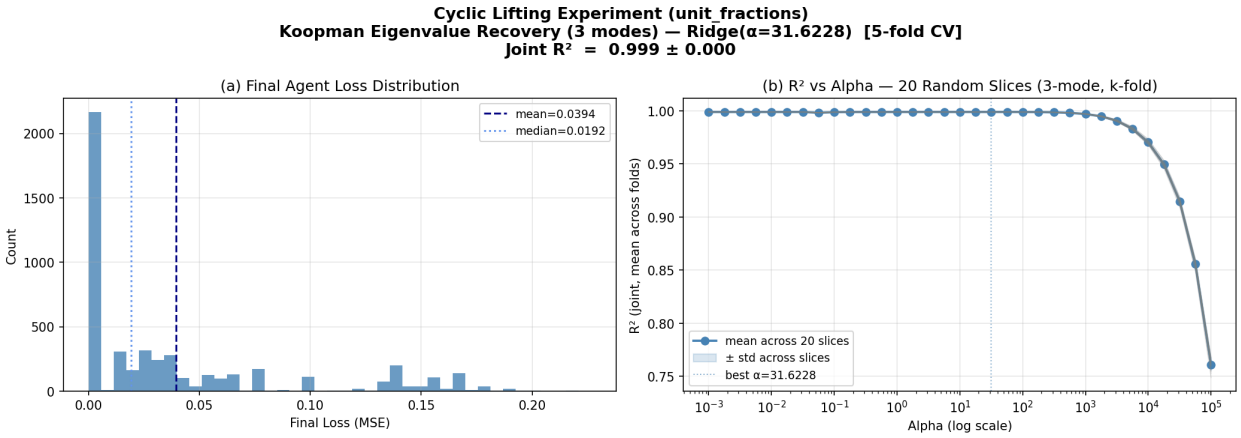

Figure 3: **Probe results, unit fractions.** *(a)* Final agent loss distribution. *(b)* Mean $R^2 \pm$ std across 5 folds vs. regularization strength α, for $P = 20$ independently drawn random parameter slices (grey curves) and their mean (blue). The curves are indistinguishable at $R^2 \approx 1.000$ across the full plateau from 10^{-3} to 10^{3}, confirming a maximally stable invariant uniformly distributed throughout the parameter vector. $N = 5{,}000$, $T = 200$, $N_{\text{modes}} = 3$, $d = 734$ from $D = 3{,}670$.

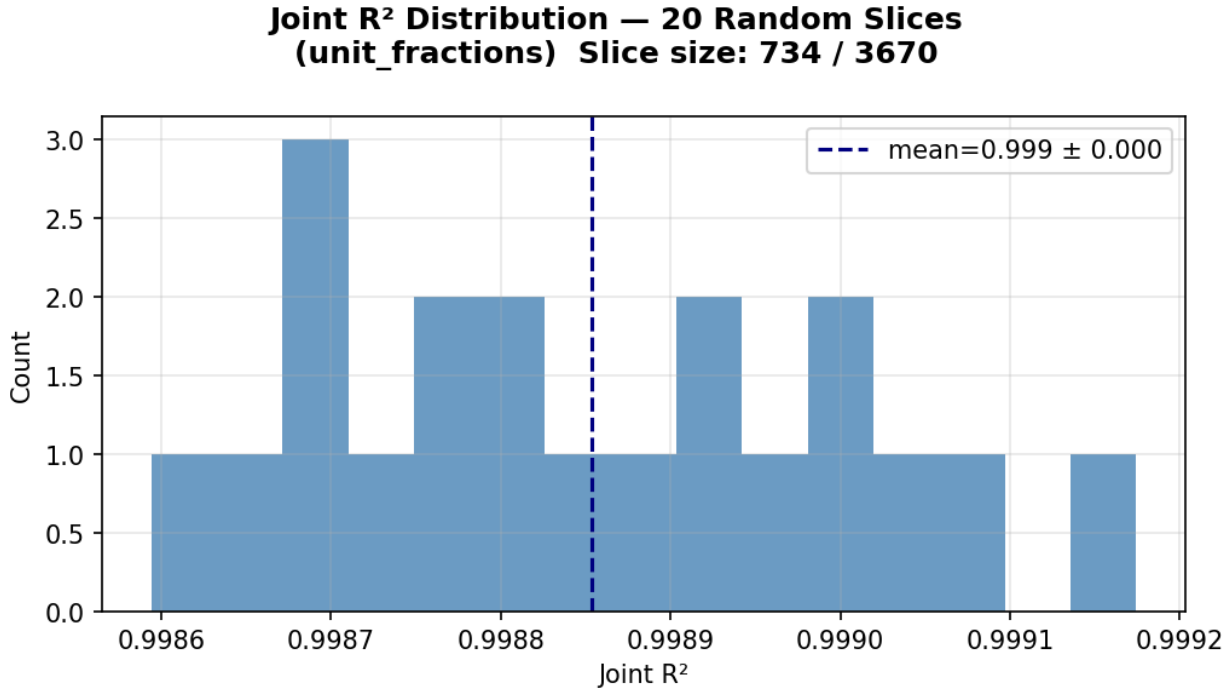

Figure 4: **Distribution of Joint R^2 across 20 random slices, unit fractions.** All 20 draws fall in the range 0.9986–0.9992 with std = 0.000, confirming that the Koopman encoding saturates any same-size window into the parameter vector. Note that the near-perfect R^2 partly reflects the discrete structure of the environment pool: twelve well-separated targets in a 734-dimensional probe space are readily regressible. The continuous-parameter random walk experiment provides the more demanding test.

entirely to the presence of Koopman structure in the unit-fraction sequences and its absence in π. The distribution of Joint R^2 across the 20 random slices (Figure 6) mirrors the positive results in its uniformity: just as structure is recoverable from any window into the parameter vector when it exists, it is absent from every window when it does not. The uniformity of the null across all 20 draws — std = 0.000 — confirms that the absence of structure is as globally distributed as its presence.

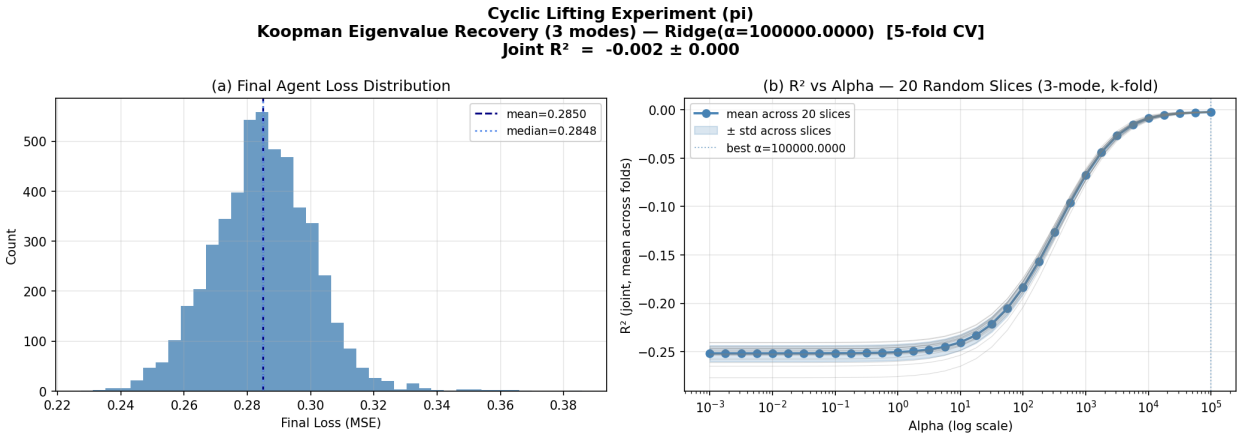

Figure 5: **Probe results, π null experiment (paired design).** *(a)* Symmetric loss distribution confirms convergence to the mean predictor. *(b)* R^2-vs-α curves for $P = 20$ random parameter slices (grey) and their mean (blue). No slice plateaus; all rise monotonically toward zero, with best $\alpha = 100{,}000$ in every case. Probe targets are identical to Figure 3; the difference is attributable entirely to the absence of spectral structure in π. Joint $R^2 = -0.002 \pm 0.000$ across all 20 slices.

11.5 Results: Nearest-Neighbor Random Walk

$N = 5{,}000$ environments, $q_i \sim \text{Uniform}(0.01, 0.99)$, $T = 200$, $N_{\text{epochs}} = 100$, $N_{\text{modes}} = 3$, single initialization, $P = 20$ random slices.

All 20 random slices yield Joint $R^2 = 0.981 \pm 0.001$, with a range of only 0.979–0.984. This is a more stringent test than the fraction expansions: a continuously varying asymmetry parameter, a stochastic observation source, and no discreteness structure that could inflate the probe.

12 Theoretical Implications

12.1 The Persistent Structural Parameter

In the quasi-periodic experiment, the persistent structural parameter is the rotation rate ω. In the unit-fraction experiment, it is the period p. In the stochastic experiment, it is the asymmetry q. In each case, the agent's converged parameters encode this parameter from prediction error alone, and a linear probe recovers it across agents sharing a common initialization. The architecture is indifferent to whether the parameter indexes a rotation rate, a period, or a stochastic drift: what matters is that it is persistent across the trajectory and distinguishes environments from one another. The use of fresh trajectories at each epoch means that trajectory-specific content cannot be encoded in the weights; what persists across training is the environment's generative constant, not any particular sequence of observations.

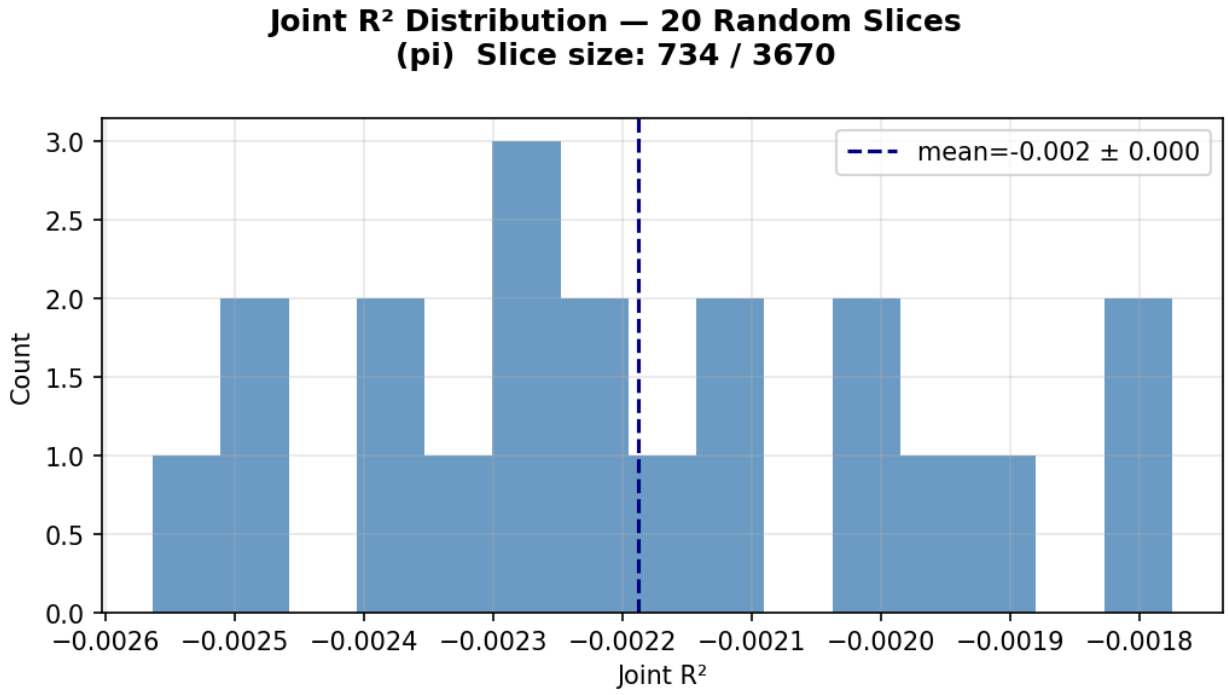

Figure 6: **Distribution of Joint R^2 across 20 random slices, π null experiment.** All 20 draws fall in the range -0.0026 to -0.0018 with std $= 0.000$. The absence of spectral structure is as uniformly distributed as its presence in the positive experiments: no slice finds anything, and the spread across draws is indistinguishable from zero.

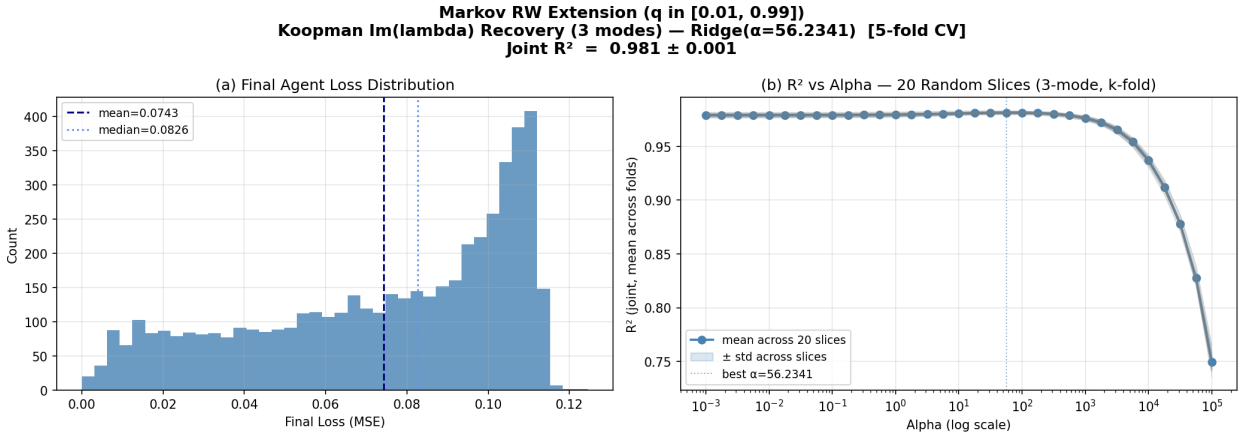

Figure 7: **Probe results, nearest-neighbor random walk.** *(a)* Final agent loss distribution. *(b)* R^2-vs-α curves for $P = 20$ random parameter slices (grey) and their mean (blue). The curves are indistinguishable throughout the plateau at $R^2 \approx 0.981$, confirming uniform distributed encoding. This is the most demanding test: 5,000 continuously varying environments, a stochastic source, and no discreteness artifact. Joint $R^2 = 0.981 \pm 0.001$, range 0.979–0.984. $N = 5{,}000$, $T = 200$, $N_{\text{modes}} = 3$, $d = 734$ from $D = 3{,}670$.

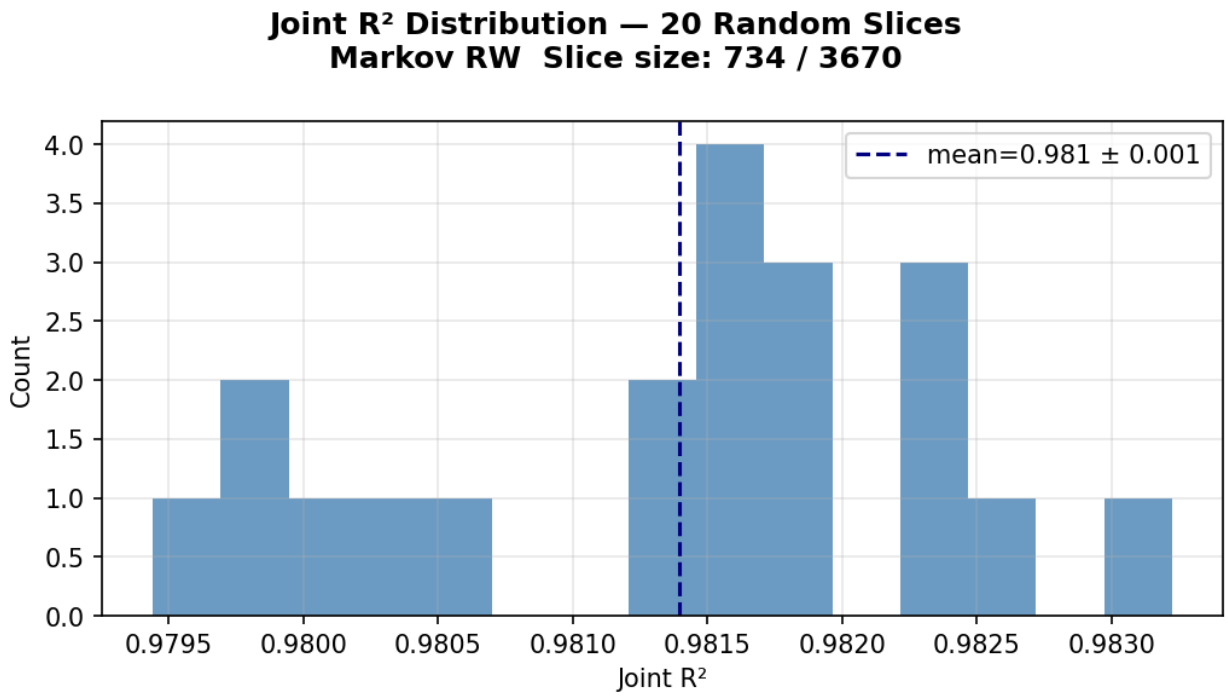

Figure 8: **Distribution of Joint R^2 across 20 random slices, nearest-neighbor random walk.** All 20 draws fall in the range 0.979–0.984 with std = 0.001. The continuously varying asymmetry q across 5,000 distinct environments is linearly readable from any same-size window into the converged parameter vector, with no discreteness artifact inflating the result.

12.2 Distributed Encoding and the Global Solution Geometry

The Koopman encoding is not localized to any privileged layer or component of the parameter vector. Across all three experiment types, $P = 20$ random slices of size $d = 734$ drawn uniformly from the full $D = 3{,}670$-dimensional parameter vector yield near-identical Joint R^2 values with standard deviation at or below 0.003: quasi-periodic 0.968 ± 0.003 (range 0.963–0.974), unit fractions 0.999 ± 0.000 (range 0.999–0.999), nearest-neighbor random walk 0.981 ± 0.001 (range 0.979–0.984), and π null -0.002 ± 0.000 (range -0.003 to -0.002). The null result is as uniformly distributed as the positive results, confirming that the probe is tracking spectral structure rather than a generic property of shared initialization.

This establishes that next-step prediction training in the relational observable space organizes the *full* parameter vector as a linear function of the Koopman spectrum of the environment. The solution that SGD converges to, within the fixed initialization coordinate frame, varies with the environment in a way that is globally and linearly consistent with the Koopman spectrum — throughout parameter space, not at any particular location within it.

12.3 Geometric Consistency and Initialization

All agents begin from identical weight vectors derived from a fixed random seed; any difference in the converged parameters is attributable solely to the environment. A single linear map trained on the agent population then recovers each environment's Koopman spectrum from the corresponding agent's parameter slice. This consistency is relative to the initialization: agents sharing a common seed converge to parameter vectors whose inter-environment variation tracks the Koopman spectrum along consistent linear directions. Agents initialized from independent random seeds do not exhibit the same cross-environment geometric consistency. A given initialization and set of hyperparameters establishes a constraint on the empirical hypothesis space of the agent, within which the environment forces linearly encodable structure.

12.4 Initialization Independence and Ensemble Averaging

The results reported above are established within a fixed initialization coordinate frame: all agents share a common weight initialization, and the inter-instance linear map is trained relative to that frame. A natural question is whether the distributed Koopman encoding is an artifact of this particular initialization or a stable property of what prediction training does to the parameter vector regardless of where it starts.

To test this, we train ensembles of $M = 15$ independently initialized agents on each of $N = 5{,}000$ random walk environments, using a common fixed set of M initialization seeds applied uniformly across all environments, and mean-pool the converged parameter vectors before probing. Initialization-specific structure is not consistent across seeds and cancels under averaging; linear Koopman structure is forced by the environment and survives. The pooled parameter vectors are passed to the same linear probe in place of individual vectors; probe targets and evaluation protocol are unchanged.

The result is Joint $R^2 = 0.981 \pm 0.002$ across $P = 20$ random slices — statistically indistinguishable from the single-initialization result of 0.981 ± 0.001. The linearity of spectral encoding is maintained under ensemble averaging, consistent with a distributed linear representation of Koopman structure being a stable property of prediction training under the present regime rather than a coordinate artifact of any particular initialization.

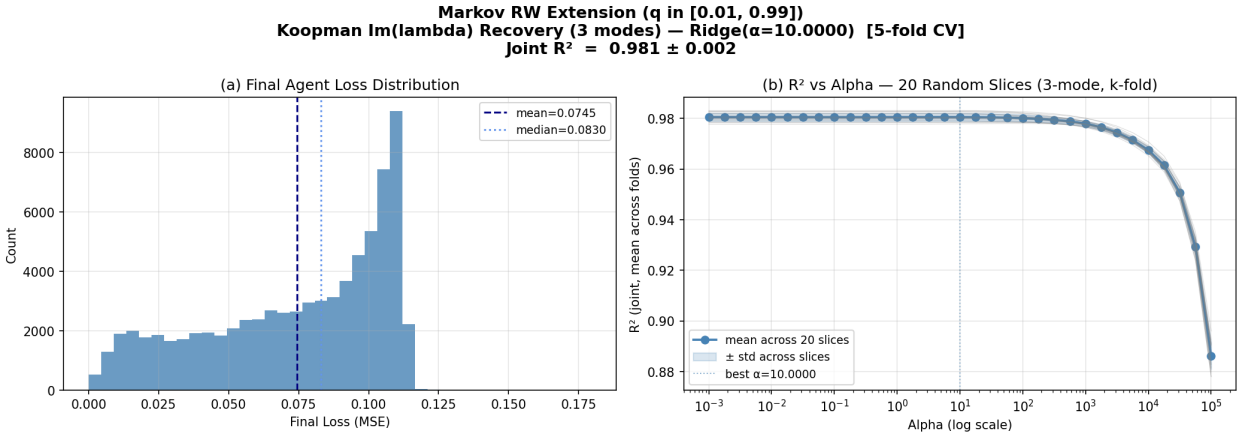

Figure 9: **Probe results, nearest-neighbor random walk, ensemble variant ($M = 15$ initializations).** *(a)* Final agent loss distribution. *(b)* R^2-vs-α curves for $P = 20$ random parameter slices (grey) and their mean (blue). The plateau is maintained at $R^2 \approx 0.981$, statistically indistinguishable from the single-initialization result, confirming that distributed linear Koopman encoding is not an artifact of any particular initialization. $N = 5{,}000$, $M = 15$, $T = 200$, $N_{\text{modes}} = 3$, $d = 734$ from $D = 3{,}670$.

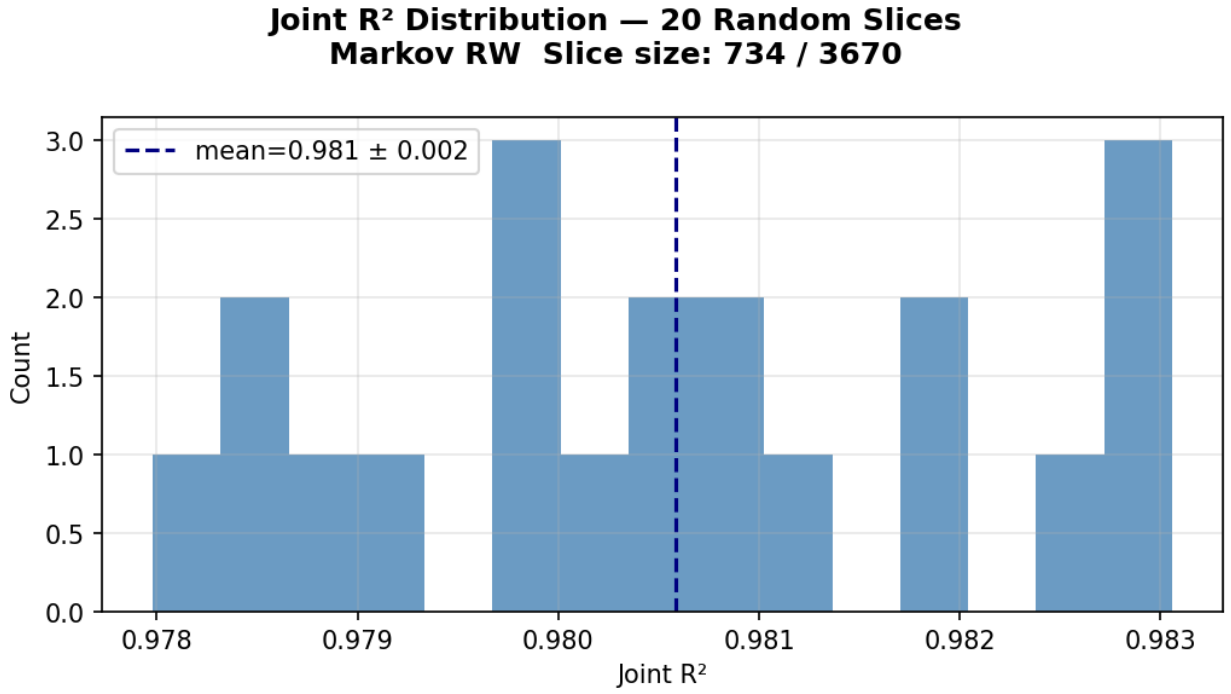

Figure 10: **Distribution of Joint R^2 across 20 random slices, nearest-neighbor random walk, ensemble variant ($M = 15$ initializations).** The distribution is tightly concentrated, with mean 0.981 ± 0.002, matching the single-initialization result within measurement error.

12.5 Relation to Data-Driven Koopman Methods

Data-driven Koopman methods [3, 4, 5] pursue learned Koopman embeddings through explicit operator and reconstruction losses. Our setting differs in two ways. First, the agent is trained only on prediction error; linearity of the dynamics in the learned coordinates is not explicitly enforced. Second, we probe the converged parameter vector directly rather than learning an embedding. The result is that linear decodability emerges as a byproduct of prediction training under the specified architecture, rather than as a directly optimized objective. The classical spectral estimation literature (DMD, Prony) can recover Koopman eigenvalues from the lifted observations in closed form with no training [2]. Our contribution is orthogonal: SGD under the specified architecture produces a global parameter geometry aligned with the Koopman spectrum in a linearly readable way, within a fixed initialization coordinate frame.

12.6 Future Directions

A natural direction is to characterize the relationship between architecture and the distributed Koopman encoding more precisely. The present architecture compresses the concatenated state–observation input of dimension $d_s + d_o = 38$ to $d_h = 32$, and this compression appears sufficient for strong global Koopman encoding, possibly by acting as a regularizer that discourages high-dimensional nonlinear solutions. The relationship between hidden layer width, training loss, and linear recoverability of the Koopman spectrum remains to be fully characterized.

A second direction is to extend the mechanism to agents embedded in continuously evolving environments, where state updates are governed by differential equations rather than discrete maps. The Koopman framework extends naturally to continuous-time dynamics [2], and the question of whether a continuous-time analogue of the present architecture would encode the continuous Koopman spectrum in a globally linear and distributed form is an open one.

A third direction concerns non-stationary environments. The classical Koopman framework assumes a stationary measure-preserving dynamical system, and data-driven spectral estimation methods inherit this assumption. The present mechanism makes a weaker demand: that the environment have enough persistent spectral structure across the training distribution for prediction error to organize it into the parameter vector. This suggests that agents trained on slowly drifting environments — where frequencies shift, asymmetry changes, or the generating distribution evolves gradually — may encode a continuously shifting Koopman characterization. As an agent trains over an arbitrary non-stationary environment, meaningful weight slices may be extracted at successive time points that classical methods, which require stationarity or explicit non-stationarity modeling, cannot produce. Whether the distributed encoding found here extends to such settings, and what a linear probe recovers if it does, is an open question.

13 Conclusion

We have shown that a predictive agent trained on prediction error alone, with the specified architecture, encodes the Koopman spectral structure of its environment in a linearly decodable form distributed throughout its converged parameter vector — across qualitatively distinct environment classes and under a uniform prior over environments.

In the quasi-periodic experiment, both independent Koopman eigenvalues are linearly recoverable with mean Joint $R^2 = 0.968 \pm 0.003$ across $P = 20$ independent random parameter slices, from $N = 5{,}000$ agents sampled from Dirichlet$(1, 1, 1)$ — the maximum-entropy distribution over the full space of quasi-periodic environments of this form. The range across slices (0.963–0.974)

and near-zero standard deviation confirm that the encoding is a global property of the solution geometry rather than a feature of any particular location in the parameter vector.

Via the canonical cyclic lifting, the same architecture and the same mechanism extend to arbitrary sequences over finite dictionaries. Unit fractions provide a deterministic positive result (Joint $R^2 = 0.999 \pm 0.000$, three modes, range 0.999–0.999 across 20 slices); the decimal expansion of π provides a paired null (Joint $R^2 = -0.002 \pm 0.000$, uniformly across all 20 slices) isolating spectral structure as the operative variable. The nearest-neighbor random walk, with continuously varying asymmetry q across 5,000 distinct environments, provides the most demanding test: Joint $R^2 = 0.981 \pm 0.001$ with a range of only 0.979–0.984 across 20 random parameter slices, confirming genuine regression with no discreteness artifact.

The inter-instance probe is geometrically consistent within a shared initialization: agents beginning from the same weight vector converge to parameter vectors whose inter-environment variation tracks the Koopman spectrum along consistent linear directions throughout. Next-step prediction training in the relational observable space organizes the full parameter vector as a linear function of the Koopman spectrum of the environment — and the linear probe recovers it from any same-size window into that vector.

References

[1] B. O. Koopman. Hamiltonian systems and transformation in Hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931.

[2] I. Mezić. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41(1):309–325, 2005.

[3] B. Lusch, J. N. Kutz, and S. L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9(1):4950, 2018.

[4] N. Takeishi, Y. Kawahara, and T. Yairi. Learning Koopman invariant subspaces for dynamic mode decomposition. In *Advances in Neural Information Processing Systems*, 2017.

[5] Q. Li, F. Dietrich, E. M. Bollt, and I. G. Kevrekidis. Extended dynamic mode decomposition with dictionary learning: a data-driven adaptive spectral decomposition of the Koopman operator. *Chaos*, 27(10):103111, 2017.

[6] J. Hewitt and C. D. Manning. Structural probes for finding syntax in word representations. In *Proceedings of NAACL-HLT*, 2019.

[7] Y. Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

[8] D. D. Wall. Normal numbers. Ph.D. thesis, University of California, Berkeley, 1949.

The Imagination Machine V: On Abstraction and Analogy

Mark Tracy
Boston University
mrktracy@bu.edu

1 Overview

Analogy is the bedrock of communication. Even that sentence makes use of analogy: as bedrock underlies and supports structures, so too does analogy underlie and support communication, allowing us to coordinate activity and manipulate our environment. Analogy allows a reasoner to transfer previously learned structure to a new situation, generating hypotheses and thereby facilitating new understanding. So fundamental is analogy to language that it proves challenging to articulate the abstract structure of analogy and to codify valid analogical reasoning. Nonetheless, it remains a fundamental endeavor for any interested in understanding mentation. In the foregoing, I introduce and augment one popular model of analogy, and I utilize the formalism thus achieved to attempt a definition of valid analogical reasoning.

2 Classical Theories of Analogy

A domain may be defined as a tuple:[1]

$$D = (O, A, R, S, T)$$

- O = set of objects
- A = set of attributes (unary operators: $a \in A \implies a : O \to S$)
- R = set of relations (n-ary operators: $r \in R \implies \exists\, n \in \mathbb{N}, \quad r : O^n \to S$)
- S = set of statements
- T = set of statements believed to be true (belief set)

Note that attributes are a special case of relations: each $a \in A$ is simply a unary relation, so formally $A \subseteq R$.

[1] This definition follows the standard treatment of domains in analogy and relational reasoning literature (cf. 1), but extends it to include a set of statements S and a belief set T, corresponding respectively to the expressible and the held-to-be-true propositions within the domain.

2.1 Structure-Mapping Theory of Analogy

In the landmark paper "Structure-Mapping: A Theoretical Framework for Analogy," Gentner argues that an analogy is a mapping between objects in a base domain and objects in a target domain that does not necessarily carry over object-level attributes but which carries over some relational predicates.[1]

2.2 A Formal Definition of Analogy

An analogy between a source domain $D_s = (O_s, A_s, R_s, S_s, T_s)$ and a target domain $D_t = (O_t, A_t, R_t, S_t, T_t)$ is defined by a tuple:

$$A = (X, Y, M, P)$$

- $X \subset O_s$: a collection of objects in the source domain
- $Y \subset O_t$: a collection of objects in the target domain
- $M : X \to Y$: a mapping of objects from source to target domain
- $P \subset \{r \mid r \in R_s \cap R_t \text{ and } \exists \mathbf{x} \in X^k \text{ for some } k \in \mathbb{N} \text{ such that } r(\mathbf{x}) \in T_s \text{ and } r(M(\mathbf{x})) \in T_t\}$: a set of relations that are present in the source and target domains, are true of some tuple of objects in the source domain, and are preserved in the target domain via the mapping M. As a notational convention, we consider $M(\mathbf{x})$ to be the component-wise application of the mapping M to the tuple $\mathbf{x}$, i.e. $\mathbf{x} = (x_1, \ldots, x_n) \implies M(\mathbf{x}) = (M(x_1), \ldots, M(x_n))$.

2.3 Analogical Reasoning

Let D_s be a source domain and D_t a target domain. Suppose:

- $X_1 \subset O_s$ is a subset of objects in the source domain. Let $|X_1| = n$.
- $Y_1 \subset O_t$ is a subset of objects in the target domain.
- $M : X_1 \to Y_1$ is a mapping of the source domain subset to the target domain subset.
- P is a set of relations preserved by the mapping M.

This establishes an analogy between D_s and D_t. Now suppose that some further fact (of a particular form to be specified below) holds in the source domain; we formally define an **analogical reasoning step** to be the positing of a corresponding form of further fact in the target domain. Formally:

Suppose there exists a superset of X_1 called X_0:

$$\begin{aligned} X_1 &\subseteq X_0 \\ |X_0| &= m \geq n \end{aligned}$$

and suppose that

$$r(\mathbf{x}^*) \in T_s$$

for some tuple $\mathbf{x}^* \in X_0^k$ for some $k \in \mathbb{N}$ and for some relational predicate $r \in (R_s \cap R_t)$.

Then an analogical reasoning step is to hypothesize that there exists a mapping M' that preserves and extends the original analogical mapping M and preserves the further observed relation in the source domain, r. In particular, the hypothesis is as follows:

$$\begin{aligned} &\exists\, Y_2 \subset O_t \quad \text{and} \\ &\exists\, M' : X_0 \quad \rightarrow \quad Y_1 \cup Y_2 \quad \text{such that} \\ &M'(x) \quad = \quad M(x) \quad \forall x \in X_1 \quad \text{and} \\ &r(M'(\mathbf{x}^*)) \quad \in \quad T_t, \end{aligned}$$

where $M'(\mathbf{x}^*)$ is the component-wise application of the mapping M' to the tuple $\mathbf{x}^*$ identified above.

This formulation captures the logic of projecting relational structures from the source domain into the target domain, conditioned on preserved analogical structure. It highlights how analogy can support hypothesizing about unseen objects, roles, or relations in the target domain by structurally mapping known relations in the source.

2.4 Analogy as Mediated by Abstraction

Abstraction, in the broadest sense, refers to the process or result of mapping a collection of objects, attributes, or relations to a single representation, typically to retain only information which is relevant for a particular purpose.

There is a connection between abstraction and analogy that is insufficiently explored in Genter's 1983 paper. If, as Gentner convincingly argues, an analogy is a mapping between objects in a base domain and objects in a target domain that does not necessarily carry over object-level attributes but which carries over some relational predicates [1], then for any analogy there exists an abstract domain that implicitly mediates the analogy. In particular, the domain that mediates an analogy $A = (X, Y, M, P)$ between a source domain $D_s = (O_s, A_s, R_s, S_s, T_s)$ and a target domain $D_t = (O_t, A_t, R_t, S_t, T_t)$ consists of:

- **A new set of objects, O_{abs}:**
  - Call them symbols.
  - $\forall x \in X, (x, M(x)) \in O_{\text{abs}}$.
  - Notational convention: for a $k-$tuple of objects in the source domain, $\mathbf{x} \in X^k$, we denote the corresponding tuple of symbols as $(\mathbf{x}, M(\mathbf{x})) \in O_{\text{abs}}^k$, where $M(\mathbf{x})$ is the component-wise application of M to $\mathbf{x}$. In particular:

$$\begin{aligned} \mathbf{x} &= (x_1, \ldots, x_k) \implies \\ M(\mathbf{x}) &= (M(x_1), \ldots, M(x_k)) \text{ and} \\ (\mathbf{x}, M(\mathbf{x})) &= ((x_1, M(x_1)), \ldots, (x_k, M(x_k))) \end{aligned}$$

- **A set of predicate attributes, A_{abs}:**

- $A_{\text{abs}} = P \cap A_s$
- The set of unary relations preserved by the analogy, if any.

- **A set of predicate relations, R_{abs}:**
  - Call them abstract relations.
  - $r \in P \iff r \in R_{\text{abs}}$
  - $r(\mathbf{x}) \in T_s$ for some $\mathbf{x} \in X^k$ with $k \in \mathbb{N} \implies r((\mathbf{x}, M(\mathbf{x}))) \in T_{\text{abs}}$.

- **A statement set, S_{abs}:**
  - All possible combinations from the collections of objects, attributes, and relations specified above.

- **A belief set, T_{abs}**
  - A subset of S_{abs}, populated as specified above.

2.4.1 An example

Take the analogy, "An atomic nucleus is like the solar system." [1] At an earlier point in scientific history, the analogical mapping may have looked like this:

$$M : X \to Y$$
$$\text{NUCLEUS} \mapsto \text{SUN}$$
$$\text{ELECTRON} \mapsto \text{PLANET}$$

And the relationships preserved include:

$$\{\text{ORBITS}, \text{IS\_MOVING}\} \subset P.$$

Now, in recognizing a mediating abstract domain we may synthesize new symbols with carried-over attributes and abstract relations, thereby forming a mediating abstract domain that both source and target instantiate:

$$\{\text{NUCLEUS}, \text{SUN}\} \mapsto \text{CENTRAL\_BODY}$$
$$\{\text{ELECTRON}, \text{PLANET}\} \mapsto \text{SATELLITE}$$
$$\text{ORBITS} \in R_{\text{abs}}$$
$$\text{IS\_MOVING} \in A_{\text{abs}} \subset R_{\text{abs}}$$

Now, obviously each instance of SATELLITE and of CENTRAL_BODY in the two original domains has attributes (mass, charge, etc.) whose values determine how the abstract relation

$$\text{ORBITS}(\text{SATELLITE}, \text{CENTRAL\_BODY})$$

manifests in these two distinct domains. Note that the statement IS_MOVING(SATELLITE) $\in S_{\text{abs}}$ happens to carry over into the belief set of this abstract domain, T_{abs}, since relative motion is characteristic of a classical satellite in both original domains.

Analogy is not simply recognizing, "D_s is like D_t". Instead, analogy is mediated by abstraction: it is to say, "D_s is like D_t because there exists an abstract domain D_{abs} of which both D_s and D_t are instances." Or, in other words, to recognize an analogy is to say, "This pattern of relations in D_s is like that pattern of relations in D_t—and there's a higher-order domain D_{abs} that generalizes both."

References

[1] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983. doi: 10.1207/s15516709cog0702_3.

The Imagination Machine VI: Holons, Horn Fillings, and the Self-Demonstration of Analogy

Mark Tracy
Boston University
mrktracy@bu.edu

Salash Tolan Nabaala

Abstract

Several frameworks arising in philosophy, mathematics, and epistemology exhibit a common structural pattern: a partially specified relational configuration is extended into a coherent higher-order structure that asymmetrically contains its constituents and may itself participate in further extensions. This paper identifies this pattern—the *extension schema*—across three primary frameworks: holonic composition, simplicial horn filling, and analogical abstraction, with a related formulation in horn-filling classification.

We demonstrate, in the native formal language of each framework, that each instantiates the schema and that the comparison between them produces an abstract mediating domain in which their shared structure becomes explicit.

The central claim is that the construction establishing this correspondence instantiates the schema itself. The holonic and simplicial frameworks together form a partially specified relational configuration, and the abstract domain that unifies them arises through the same extension operation the schema describes. The argument therefore exhibits the structure it analyzes: the reader witnesses the schema execute in the course of the proof.

1 Introduction

In many mathematical and conceptual settings, coherent structures arise by extending partially specified relational configurations. Some collection of objects and relations determines most of the structure of a larger whole, but one higher-order relational element remains unspecified. An extension operation produces a coherent unity that contains the original configuration as a proper part, is not reducible to it, and may itself participate in further constructions of the same kind.

This paper identifies a common instance of this pattern—the *extension schema*—across three frameworks: the metaphysical notion of holons [4], the mathematical operation of horn filling in simplicial sets, and the construction of abstract mediating domains in analogical reasoning [2]. The aim is not to claim that these frameworks describe the same objects in any literal sense. It is to show, in the language of each formalism, that each is a genuine instantiation of the same abstract structural pattern, and that the act of showing this is itself a further instantiation.

The paper proceeds as follows. Sections 2 through 4 introduce the three frameworks. Section 5 states the extension schema and proves that each framework instantiates it, with a separate demonstration in the native language of each formalism. Section 6 shows that the construction performed in Section 5 is itself a fourth instantiation, occurring as the reader follows the argument. Section 7 discusses the recursive structure common to all three frameworks. Section 8 concludes.

2 Analogy as Mediated by Abstraction

Definition 1 (Domain)**.** *A domain is a tuple $D = (O, A, R, S, T)$ where O is a set of objects; A is a set of attributes (unary relations $a : O \to S$); R is a set of relations (each $r \in R$ an n-ary map $r : O^n \to S$ for some $n \in \mathbb{N}$); S is a set of statements; and $T \subseteq S$ is a set of accepted statements. Since every attribute is a unary relation, $A \subseteq R$.*

Definition 2 (Analogy)**.** *An analogy between domains $D_s = (O_s, A_s, R_s, S_s, T_s)$ and $D_t = (O_t, A_t, R_t, S_t, T_t)$ is a tuple $\mathcal{A} = (X, Y, M, P)$ where $X \subseteq O_s$, $Y \subseteq O_t$, $M : X \to Y$ is a mapping of objects, and $P \subseteq R_s \cap R_t$ is a set of relations preserved by M: for each $r \in P$ and tuple $x = (x_1, \ldots, x_k) \in X^k$, if $r(x) \in T_s$ then $r(M(x)) \in T_t$, where M is applied component-wise: $M(x) = (M(x_1), \ldots, M(x_k))$.*

Definition 3 (Abstract Mediating Domain)**.** *Given an analogy $\mathcal{A} = (X, Y, M, P)$ between D_s and D_t, the abstract mediating domain $D_{\text{abs}} = (O_{\text{abs}}, A_{\text{abs}}, R_{\text{abs}}, S_{\text{abs}}, T_{\text{abs}})$ is defined by:*

(i) $O_{\text{abs}} = \{(x, M(x)) \mid x \in X\}$*, whose elements are called* symbols*; for a tuple $x = (x_1, \ldots, x_k) \in X^k$, the corresponding tuple of symbols is $((x_1, M(x_1)), \ldots, (x_k, M(x_k)))$;*

(ii) $A_{\text{abs}} = P \cap A_s$*, the unary relations preserved by the analogy, called* abstract attributes*;*

(iii) $R_{\text{abs}} = P$*, called* abstract relations*;*

(iv) S_{abs} *consists of all statements expressible from O_{abs}, A_{abs}, and R_{abs};*

(v) T_{abs} *contains $r((x_1, M(x_1)), \ldots, (x_k, M(x_k)))$ whenever $r(x_1, \ldots, x_k) \in T_s$ for $r \in P$.*

The canonical projections $\pi_s(x, M(x)) = x$ and $\pi_t(x, M(x)) = M(x)$ exhibit D_s and D_t as instantiations of D_{abs}.

Remark 1. *The symbols in O_{abs} belong to neither D_s nor D_t; they encode the correspondence itself. D_{abs} is a genuinely new domain, not reducible to either source or target, and both source and target are recoverable from it by projection.*

Definition 4 (Analogical Reasoning Step)**.** *Given $\mathcal{A} = (X, Y, M, P)$ and a superset $X_0 \supseteq X$, suppose $r \in R_s \cap R_t$ and $r(x^*) \in T_s$ for some tuple $x^* \in X_0^k$. An analogical reasoning step hypothesizes the existence of a set $Y_2 \subseteq O_t$ of additional target objects and an extension $M' : X_0 \to Y \cup Y_2$ of M such that $M'(x) = M(x)$ for all $x \in X$ and $r(M'(x^*)) \in T_t$, where M' is applied component-wise to the tuple x^*. Known relational structure in the source domain licenses the projection of new structure into the target, conditioned on the preserved relational pattern.*

3 Holons

Definition 5 (Holon)**.** *A holon is an entity H such that: (i) H forms a coherent unit; (ii) H has proper parts; (iii) H may itself occur as a part of a larger entity; (iv) relations between H and its parts are asymmetric.*

Definition 6 (Holonic Containment)**.** *Write $B \prec A$ if B is a proper part of A and A contains relational structure not present in B alone. The relation $\prec$ is irreflexive and asymmetric.*

Definition 7 (Holonic Completion)**.** *Given entities $\mathcal{F} = \{B_1, \ldots, B_m\}$ with relational structure $\mathcal{R}$ among them, a holonic completion is an entity H such that: (i) $B_i \prec H$ for all i; (ii) H unifies the B_i into a coherent whole; (iii) H is not reducible to any proper subset of $\mathcal{F}$.*

Definition 8 (Holonic Hierarchy)**.** *A holonic hierarchy is a sequence $H_0 \prec H_1 \prec H_2 \prec \cdots$ in which each entity is a holonic completion of a family drawn from the previous level.*

4 Horn Filling in Simplicial Sets

Definition 9 (Simplicial Set)**.** *A simplicial set X consists of sets X_n of n-simplices for each $n \geq 0$, together with face maps $d_i : X_n \to X_{n-1}$ and degeneracy maps $s_i : X_n \to X_{n+1}$ satisfying the simplicial identities. An n-simplex $\sigma \in X_n$ represents a coherent relational configuration among $n+1$ vertices.*

Definition 10 (Horn)**.** *For $n \geq 1$ and $0 \leq k \leq n$, the kth horn Λ^n_k is the simplicial subset of Δ^n generated by all faces $d_i\iota$ for $i \neq k$, where $\iota : \Delta^n \to \Delta^n$ is the identity map. A horn is a partially specified simplex: it contains all but one of the codimension-one faces of Δ^n, with the kth face and the interior absent.*

Definition 11 (Horn Filling)**.** *A horn filling for a map $\sigma : \Lambda^n_k \to X$ is an extension*

$$\sigma' : \Delta^n \to X$$

such that $\sigma' \circ i^n_k = \sigma$, where $i^n_k : \Lambda^n_k \hookrightarrow \Delta^n$ is the inclusion. The filled simplex $\sigma'(\iota) \in X_n$ completes the partial relational data specified by σ.

Remark 2 (Extension and lifting)**.** *Horn filling may be interpreted categorically as a lifting problem: a morphism defined on the partial simplicial object Λ^n_k extends to a morphism on the full simplex Δ^n. Partial relational data is extended to a coherent higher-dimensional simplex.*

Definition 12 (Face Containment)**.** *For simplices $\tau \in X_m$ and $\sigma \in X_n$ with $m < n$, write $\tau \prec_s \sigma$ if τ is a face of σ, that is, $\tau = d_{i_1} \cdots d_{i_j}\sigma$ for some sequence of face maps.*

5 The Extension Schema and Its Instantiations

Definition 13 (Extension Schema)**.** *An extension schema consists of:*

(i) a partially specified relational configuration C_{partial};

(ii) an extension operation ϕ producing a coherent structure $C_{\text{whole}} = \phi(C_{\text{partial}})$;

(iii) an asymmetric containment relation $C_{\text{partial}} \prec C_{\text{whole}}$: the partial configuration contributes to but does not exhaust the whole;

(iv) a recursion rule: C_{whole} may itself serve as C_{partial} in a further application of ϕ.

Theorem 1 (Structural Correspondence)**.** *Holonic composition, simplicial horn filling, and analogical abstraction each instantiate the extension schema. We demonstrate this in the native formal language of each framework.*

Proof. We treat each framework in turn, exhibiting all four components of Definition 13 explicitly.

Case 1: Holonic composition.

Partial configuration. Let $\mathcal{F} = \{B_1, \ldots, B_m\}$ be a family of entities bearing relational structure $\mathcal{R}$ among them. The pair $(\mathcal{F}, \mathcal{R})$ specifies how the constituents are related but does not yet determine any unified entity containing them. This is C_{partial} in the holonic language: a collection of parts and their mutual relations, fully specified, but not yet gathered into a whole.

Extension operation. Holonic completion (Definition 7) is ϕ. Applied to $(\mathcal{F}, \mathcal{R})$, it produces a holon H that unifies $\mathcal{F}$ under $\mathcal{R}$ into a single coherent entity. H is not a new relation among the

B_i; it is a new entity whose existence is licensed by the relational structure $\mathcal{R}$ but is not identical to it. This is C_{whole}.

Asymmetric containment. By Definition 6, each $B_i \prec H$. The holon H contains the relational structure $\mathcal{R}$ among its parts and additionally the higher-order unity that no individual B_i or proper subcollection of $\mathcal{F}$ possesses. Conversely, $H \not\prec B_i$ for any i: the whole is not a part of any of its parts. The containment is strict and asymmetric.

Recursion. The holon H satisfies Definition 5 and is therefore itself eligible to serve as a member B_j of a further family $\mathcal{F}'$. Bearing new relations $\mathcal{R}'$ to other holons, H may participate in a further holonic completion H' with $H \prec H'$. The output of one completion is the input to the next.

Case 2: Simplicial horn filling.

Partial configuration. Let $\sigma : \Lambda^n_k \to X$ be a horn map. The horn Λ^n_k contains the faces $d_i\iota$ for all $i \neq k$: every codimension-one face of a would-be n-simplex is present except the kth. All pairwise, triple, and higher-order relations among the $n+1$ vertices are specified except for the one n-ary relation encoded by the missing kth face and the interior. This is C_{partial}: a relational configuration that is almost complete but lacks exactly one higher-order coherence datum.

Extension operation. Horn filling (Definition 11) is ϕ. It produces an extension $\sigma' : \Delta^n \to X$ of σ across the inclusion $\Lambda^n_k \hookrightarrow \Delta^n$, supplying the missing kth face $d_k(\sigma'(\iota)) \in X_{n-1}$ and the interior n-simplex $\sigma'(\iota) \in X_n$. The filled simplex $\sigma'(\iota)$ is a coherent n-simplex that did not exist in X before the filling. This is C_{whole}.

Asymmetric containment. For each i, the face $d_i(\sigma'(\iota)) \in X_{n-1}$ satisfies $d_i(\sigma'(\iota)) \prec_s \sigma'(\iota)$ in the sense of Definition 12. The filled n-simplex encodes a relation among all $n+1$ vertices simultaneously, which no $(n-1)$-dimensional face encodes. Conversely, no face contains the simplex that contains it: the containment is strict, asymmetric, and dimension-raising.

Recursion. The filled simplex $\sigma'(\iota) \in X_n$ is an element of X_n and may appear as the jth face of an $(n+1)$-simplex $\tau \in X_{n+1}$, that is, $d_j(\tau) = \sigma'(\iota)$ for some j. If the horn at dimension $n+1$ whose jth face is $\sigma'(\iota)$ admits a filling, then $\sigma'(\iota) \prec_s \tau$ and horn filling at dimension n has produced the input to horn filling at dimension $n+1$. The recursion follows from the fact that filled simplices are simplices.

Case 3: Analogical abstraction.

Partial configuration. Let $\mathcal{A} = (X, Y, M, P)$ be an analogy between D_s and D_t. The pair (D_s, D_t) together with M and P constitutes a partially specified relational configuration: the shared structure P is implicit in both domains, instantiated concretely in each, but the abstract domain of which both are instances does not yet exist as an explicit object. Like a horn, the data (D_s, D_t, M, P) contains enough face information to determine a coherent higher-order structure, but that structure is absent. This is C_{partial}.

Extension operation. The construction of D_{abs} (Definition 3) is ϕ. Given (D_s, D_t, M, P), it produces a new domain whose objects are the symbols $(x, M(x))$, whose attributes are the preserved unary relations $P \cap A_s$, whose relations are the abstract relations P, and whose accepted statements are those licensed by the preserved relational structure. D_{abs} is not a subset or quotient of D_s or D_t; its objects, the symbols, exist in neither source nor target. It is a genuinely new domain. This is C_{whole}.

Asymmetric containment. The projections π_s and π_t exhibit D_s and D_t as instantiations of D_{abs}, and this form of containment is asymmetric. D_{abs} contains the symbols $(x, M(x))$ present in neither D_s nor D_t alone. Neither source nor target determines D_{abs} individually; the abstract domain requires both, together with M and P. Conversely, D_s and D_t are each instantiations of D_{abs}. Instantiation and generalization function, then, as a form of asymmetric containment: $D_s \prec D_{\text{abs}}$ and $D_t \prec D_{\text{abs}}$.

Recursion. D_{abs} satisfies Definition 1 and is itself a domain. It may serve as source or target in a further analogy $\mathcal{A}'$ with a new domain D_u, producing a further abstract mediating domain D'_{abs} of which both D_{abs} and D_u are instances, with $D_{\text{abs}} \prec D'_{\text{abs}}$. The extension operation applies again at a higher level of abstraction.

In each case all four components of the extension schema are exhibited in the native language of the framework. The schema is not imposed from outside; it is read off from the structure each framework already possesses. □

Proposition 1 (Classification as an instance of the extension schema)**.** *Let X be a simplicial set and let $f : X \to S$ be a map satisfying the following horn-extension condition: for every horn $\sigma : \Lambda^n_k \to X$ with $n \geq 2$ there exists a simplex $\sigma' : \Delta^n \to S$ such that*

$$\sigma' \circ i^n_k = f \circ \sigma.$$

Then the operation induced by f instantiates the extension schema of Definition 13.

Proof. The restriction $n \geq 2$ excludes the degenerate case $n = 1$, in which a horn Λ^1_k is a single vertex and filling it imposes no coherence constraint; the substantive extension pattern begins at dimension 2, where a horn specifies two vertices of a triangle and the filling supplies the third edge and interior.

A horn $\sigma : \Lambda^n_k \to X$ specifies a partially determined relational configuration among $n + 1$ vertices, missing exactly one face and the interior of the corresponding simplex. This is C_{partial}.

The horn-extension condition ensures the existence of a simplex $\sigma' : \Delta^n \to S$ completing this configuration. The filled simplex constitutes C_{whole}.

Containment is asymmetric: the faces of Δ^n include the original horn but encode strictly less relational structure than the full simplex. The resulting simplices may themselves participate in further horn configurations in higher dimensions, yielding recursion.

Thus classification by horn filling satisfies all four components of the extension schema. □

Remark 3 (Horn-filling classification)**.** *The interpretation of classification in terms of horn-filling conditions in simplicial sets arose in discussions with Salash Tolan Nabaala. In that formulation, an environment is modeled as a simplicial set (or more generally an ∞-category) X, and a classifier is represented by a map $f : X \to S$ satisfying a horn-extension property: whenever a horn $\sigma : \Lambda^n_k \to X$ specifies partial relational structure in the environment, there exists a coherent completion $\sigma' : \Delta^n \to S$ making the diagram commute. In this sense, classification may be understood as the completion of relational configurations under an appropriate coherence constraint.*

Iterating this idea leads naturally to a hierarchy of classifiers: classifiers of the environment, classifiers of classifiers, and so on. Such a hierarchy suggests the possibility of a stabilizing level at which further iterations introduce no essentially new structure. The horn-filling account of classification can therefore be understood as another instance of the extension schema introduced in this paper. Just as horn filling extends partial simplicial configurations to full simplices, classification extends partial relational structure in the environment to coherent representations. The categorical formulation of classification described above is due to Nabaala and provides a concrete mathematical instantiation of the more general extension principle analyzed here.

6 Self-Demonstration

The proof of Theorem 1 identifies the extension schema as the abstract structure common to the three frameworks. We now observe that this identification is itself a fourth instantiation of the schema, and that the reader has just watched it execute.

Theorem 2 (Self-Demonstration)**.** *The construction performed in Theorem 1 instantiates the extension schema.*

Proof. We exhibit the four components.

Partial configuration. Prior to Theorem 1, the holonic framework D_s and the simplicial framework D_t each implicitly instantiate the extension schema within their own formalisms. But the abstract structure they share has not been made explicit as an object. The pair (D_s, D_t) is therefore a horn: it contains two concrete faces of a higher-order coherent structure—two instantiations of the schema—but the abstract domain of which both are instances is absent. This is C_{partial}.

Extension operation. The construction of Theorem 1 is ϕ. By treating the holonic framework as source domain and the simplicial framework as target domain, constructing the mapping M between their corresponding constructs, identifying the preserved relations P as the four conditions of Definition 13, and applying Definition 3, the theorem produces D_{abs}: the extension schema itself, now explicit as a domain. This is C_{whole}.

Asymmetric containment. The extension schema D_{abs} contains the symbols encoding the correspondence between holonic and simplicial constructs, and the abstract relations that both frameworks instantiate. Neither framework alone determines it. Conversely, both frameworks are recoverable from D_{abs} by projection. Both are proper parts of the extension schema: $D_s \prec D_{\text{abs}}$ and $D_t \prec D_{\text{abs}}$.

Explicit analogy $\mathcal{A} = (X, Y, M, P)$ for Theorem 2

We make the underlying analogy explicit in the terms of Definition 2. The source domain D_s is the holonic framework and the target domain D_t is the simplicial framework.

Objects $X \subseteq O_s$ and $Y \subseteq O_t$. The three object-level schema components as they appear in each framework:

$$X = \{\ (\mathcal{F}, \mathcal{R}), \quad \phi_H, \quad \prec\ \} \qquad Y = \{\ \sigma : \Lambda^n_k \to X, \quad \phi_S, \quad \prec_s\ \}$$

The mapping $M : X \to Y$.

$$\begin{aligned} (\mathcal{F}, \mathcal{R}) &\longmapsto \sigma : \Lambda^n_k \to X && \text{(partial configuration)} \\ \phi_H &\longmapsto \phi_S && \text{(extension operation)} \\ \prec &\longmapsto \prec_s && \text{(asymmetric containment)} \end{aligned}$$

Preserved relations P and the recursion attribute. The first three conditions of Definition 13 appear as preserved relations $P \subseteq R_s \cap R_t$, and M preserves each: wherever a holonic construct instantiates one of these conditions, its image under M instantiates the same condition in the simplicial language.

The recursion rule—condition (iv)—is not a fourth object in O_{abs} but an *abstract attribute* $\rho \in A_{\text{abs}} = P \cap A_s$: a unary relation expressing that each schema component is eligible to re-enter the process as a new C_{partial}. It holds of every object in O_s (holons are holons, so each $x \in X$ satisfies $\rho(x) \in T_s$) and is preserved by M (filled simplices are simplices, so $\rho(M(x)) \in T_t$ for each $x \in X$). Accordingly, T_{abs} contains $\rho(x, M(x))$ for each symbol $(x, M(x)) \in O_{\text{abs}}$: the recursion rule is an accepted statement about each object-level symbol, not a symbol itself.

Symbols $O_{\text{abs}} = \{(x,\ M(x)) \mid x \in X\}$. The objects of D_{abs} are the three pairs:

$$\begin{aligned} &\big((\mathcal{F}, \mathcal{R}),\ \ \sigma : \Lambda^n_k \to X\big) \\ &\big(\phi_H,\ \ \phi_S\big) \\ &\big(\prec,\ \ \prec_s\big) \end{aligned}$$

These symbols belong to neither D_s nor D_t. They encode the correspondence itself. The recursion attribute ρ holds of each, so T_{abs} records that every object-level component of the schema is eligible to participate in a further extension. D_{abs}—the extension schema, now explicit as a domain—is the genuinely new object constituted by this mapping. Both frameworks are recoverable from it by the projections $\pi_s(x, M(x)) = x$ and $\pi_t(x, M(x)) = M(x)$.

Recursion. D_{abs}—the extension schema, now explicit—is itself a domain and may serve as source or target in a further analogy: for instance, with the inference-implication loop of embedded epistemic systems [1], with classifier hierarchies, or with the institutional transmission of knowledge [1]. Each such analogy would produce a new abstract mediating domain at a higher level of abstraction, with D_{abs} as a proper part of it. □

Remark 4 (The warrant of self-demonstration)**.** *The self-demonstration of Theorem 2 is the paper's primary epistemic warrant, not a secondary illustration appended to an independent argument. The correspondence between the three frameworks does not rest on an external standard of correctness applied after the fact. It rests on the fact that the construction which establishes the correspondence is the same operation the schema describes.*

This is not a vicious circularity. A vicious circle assumes its conclusion in its premises. Here, the conclusion—that the construction instantiates the schema—is established by exhibiting all four components of the schema in the construction itself, exactly as Theorem 1 establishes its conclusion by exhibiting all four components in each framework. The self-demonstration is a fixed point, not a loop: the operation applied to the pair (D_s, D_t) produces an output that is an instance of the operation itself. This is the same structure as a self-consistent world model in the sense of [1]—stability under one's own operations, rather than correspondence with an external standard.

A reader disposed to deny the correspondence would have to identify the shared relational structure between holons and simplices and abstract it into a domain of which both are instances. That act is itself an instantiation of the extension schema. The schema cannot be denied from outside, because there is no outside from which to deny it that is not already inside it.

7 Recursive Structure

The recursion rule of condition (iv) in Definition 13 is not an independent stipulation. It follows from a structural feature common to all three frameworks.

Proposition 2. *In each of the three frameworks, ϕ produces structures of the same type as the elements of C_{partial}. The recursion rule therefore requires no additional hypothesis.*

Proof. A holonic completion H satisfies Definition 5 and is therefore itself a holon, eligible to serve as a member of a further family $\mathcal{F}'$. A filled n-simplex $\sigma'(\iota)$ is an element of X_n and is therefore itself a simplex, eligible to appear as a face in a higher-dimensional simplex. An abstract mediating domain D_{abs} satisfies Definition 1 and is therefore itself a domain, eligible to serve as source or target in a further analogy. In each case the output type matches the input type, and the recursion follows. □

The paper itself enacts this recursion. The extension schema D_{abs} produced in Theorem 1 immediately serves as a constituent in Theorem 2, where it participates in a further instantiation of the schema one level up. The hierarchy has already begun by the time the reader reaches this sentence.

A closely related instance of the extension schema appears in [1]. There, a world model $w \in W$ generates an observational profile through the implication map $g : W \to \Gamma$, while the inference map $F : \Gamma \to W$ produces revised models from observational data. Their composition $T = F \circ g$ defines an operator on model space. A self-consistent world model is a fixed point $w^* \in W^*$ satisfying $T(w^*) = w^*$. From the perspective of the extension schema, a provisional model together with its observational profile forms a partially specified relational configuration; the operator T is the extension operation; and a fixed point is a completed whole that is stable under its own operations. The iterative search for fixed points is the recursive structure of the schema applied to epistemology. That framework is therefore a further instance of the same pattern, and the extension schema is the abstract mediating domain between it and the frameworks treated here.

8 Conclusion

Three frameworks—holonic composition, simplicial horn filling, and analogical abstraction—instantiate a common extension schema: the pattern by which a partially specified relational configuration is extended into a coherent structure that asymmetrically contains its constituents and may participate in further extensions. This paper has demonstrated this instantiation in the native formal language of each framework, and has shown that the demonstration is itself a fourth instantiation.

The extension schema is not a new formalism imposed on these frameworks from outside. It is the abstract mediating domain of an analogy between them, constructed by the same operation it describes. A reader who has followed the argument has not only read about the schema; they have watched it execute in three cases and participated in its fourth execution.

The recursive structure established in Proposition 2 means that this is not a terminus. The extension schema, now explicit as a domain, may be placed in analogy with further frameworks—the inference-implication loop of [1], the institutional transmission of closures in [1], or classifier hierarchies in formal language theory—generating new abstract mediating domains at higher levels of abstraction. Each such construction is a further instantiation of the pattern that produced it. The schema propagates itself forward by being what it is.

References

[1] Tracy, M. *The Imagination Machine I: A View from Somewhere*. Unpublished manuscript, Boston University.

[2] Tracy, M. *The Imagination Machine V: On Abstraction and Analogy*. Unpublished manuscript, Boston University.

[3] Gentner, D. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983.

[4] Koestler, A. *The Ghost in the Machine*. Hutchinson, 1967.

The Imagination Machine VII: A Geometric Theology of the Embedded Observer

A Personal Note on the Intuition Underlying the Series

Mark Tracy
Boston University
mrktracy@bu.edu

Abstract

This paper functions as a record of recognition of the intuition that preceded the rest of the series. Its central contribution is the argument that while embeddedness forecloses to any given agent the sort of global view necessary to make a statement about the whole in which they are embedded, the universality and necessity of such foreclosure among embedded subjects suggests one particular closed geometry as a faithful model of the shared structural relationship between the whole and its embedded participants.

Contents

1 The Geometry of Maximal Uncertainty

1.1 The Medieval Formula

A formula attributed to the *Liber XXIV Philosophorum* (c. 12th century), later associated with Pascal, Giordano Bruno, and Meister Eckhart, states:

> *God is a circle whose center is everywhere and whose circumference is nowhere.*

1.2 The Hypersphere

Let the embedded observer inhabit a three-dimensional space $\mathbb{R}^3$. A sphere in $\mathbb{R}^3$ has a center locatable at a point and a boundary at finite radius. The formula is not satisfiable within $\mathbb{R}^3$.

Add one dimension. Consider the four-dimensional hypersphere

$$S^3 = \{x \in \mathbb{R}^4 : \|x\| = r\}$$

for some radius $r > 0$. From the perspective of an observer embedded within S^3—constrained to its three-dimensional surface—the following hold:

1. **Center is everywhere.** The center of S^3 lies in the fourth dimension, inaccessible to the embedded observer. Every point on S^3 is equidistant from this center. No point within the observable manifold is the center; every point is equally proximate to it.

2. **Circumference is nowhere.** S^3 has no boundary within itself. An embedded observer moving in any direction never encounters an edge.

The formula is therefore a precise description of S^3 as encountered from within.

1.3 Maximal Uncertainty as the Warrant for the Geometry

The claim that the containing structure has the geometry of S^3 might seem like a strong assumption. It is the opposite. It is the assumption that makes the fewest additional commitments beyond closure and what embeddedness itself implies.

An embedded observer—one with no access to an external vantage point, which is the founding constraint of the entire series—cannot in principle determine the global geometry of the structure it inhabits. Local measurements are consistent with many global topologies.

The question is therefore not which geometry is correct, but which geometry should be assumed in the absence of information that embeddedness itself renders inaccessible.

The hypersphere S^3 is the answer to that question. It is, among closed three-manifolds, the geometry of maximal symmetry: every point is equivalent to every other, no direction is distinguished, no boundary is present, and no center is locatable from within. To assume S^3 is to assume nothing about which region of the containing structure one inhabits, nothing about preferred directions, nothing about partitions of paths through the universe, and nothing about edges or limits. Any other closed geometry breaks at least one of these symmetries and thereby assumes more than the embedded observer can know.

Maximal epistemic humility about the global structure—the stance the framework demands of any embedded epistemic system—selects S^3 uniquely among the candidate geometries, if one is to demand closure or attempt any description at all. The medieval formula is neither an inspired guess, not merely a didactic tool. It is an epistemically honest attempt by an embedded observer to describe the structure that contains and pervades it.

1.4 The Containing Structure

The theological claim is not that God resembles a hypersphere. It is that the containing structure of being—what the series calls Ω, the universe treated as a single relational structure—bears a relationship to the embedded observer as the geometry of S^3 to three-dimensional cross-sections within its surface volume.

This is continuous with the block universe framing of *The Imagination Machine I* [1]. The universe Ω is treated there as a static relational structure containing observations, models, and consistency relations simultaneously. The atemporal character of Ω corresponds naturally to the geometry of S^3: there is no privileged temporal direction in the containing manifold, only the experience of time as the projection of four-dimensional structure onto the three-dimensional observational profile of an embedded system.

2 The Trinitarian Structure

2.1 A Triad from the Geometry

Let $\mathcal{B}$ denote the four-dimensional containing structure (the hypersphere S^3 as living whole). Let $\mathcal{E}$ denote a three-dimensional cross-section of $\mathcal{B}$. The embedding relation is the map

$$\iota : \mathcal{E} \hookrightarrow \mathcal{B}$$

This gives a natural triad:

$$(\mathcal{B},\ \mathcal{E},\ \iota)$$

a four-dimensional whole, its three-dimensional expression, and the dynamic relation between them.

3 The Fixed Point and Its Theological Register

3.1 The Inference–Implication Loop

The formal structure of *The Imagination Machine I* [1] is the inference–implication loop:

$$\Gamma \xrightarrow{F} W \xrightarrow{g} \Gamma$$

with induced operator $T = F \circ g : W \to W$. A self-consistent world model is a fixed point:

$$T(w^*) = w^*$$

From the geometric perspective, the fixed-point condition is the formal expression of what it means for a three-dimensional cross-section to correctly reflect the four-dimensional containing structure: a model whose implied observational profile, when resubmitted to inference, reproduces itself.

3.2 Calibration as Orientation

The measure μ_D over the observation space represents the empirical distribution of observations induced by the geometry of Ω. Calibration—the alignment between a system's empirical trajectory and the global observational distribution—is, in this register, the alignment of the observer's internal model with the structure of what contains it.

Miscalibration is a form of ontological disorientation: the observer's predictions, while locally stable, diverge from the structure of what contains it. The three failure modes of *The Imagination Machine I* [1]—dogmatism, miscalibration, and the irreducibility of will—correspond to three modes of estrangement: refusal to refine or modify, a distorted image of the whole, and the irreducible freedom that persists to expose or foreclose experiences even when both are functioning correctly.

3.3 Will as the Irreducible Remainder

The Imagination Machine I [1] is explicit: the inference–implication loop determines the space of stable closures W^*, but does not determine which element of W^* is instantiated. Will is what remains when the loop has done everything it can do.

Theologically, this is the formal location of freedom. The containing structure does not determine which stable closure the embedded observer instantiates. The observer must choose, in territory no model can fully exhaust. This is the formal structure of what the tradition calls grace and response: the geometry makes the fixed point available; the instantiation is the observer's act. The framework does not resolve this. It locates it with precision, which is what a framework can do.

4 Brief Orientation to the Literature

This section locates the account within existing theological literature for readers who approach it from that direction. It is not an argument; it is a map.

The closest existing category is **panentheism**—the view that the world is contained within God without being identical to God, and that God is not exhausted by the world. The present account is panentheistic in structure: $\mathcal{E} \subset \mathcal{B}$ but $\mathcal{B} \neq \mathcal{E}$. The fourth dimension of $\mathcal{B}$ is inaccessible to the embedded observer; it is the formal location of transcendence. The difference from standard panentheism is that the containment relation here has a geometric rather than merely metaphorical expression. See Hartshorne [11] for the classical panentheist position.

The **via negativa**—associated with Pseudo-Dionysius, Meister Eckhart [13], and the *Cloud of Unknowing*—holds that God cannot be positively characterized, only approached by negation. The present account provides a formal account of why: the apprehension of $\mathcal{B}$ is not accessible to the embedded observer because $\mathcal{B}$ comprises its own comprehension. The apophatic tradition is the recognition, in the vocabulary available to it, of this geometric inaccessibility. Negative theology is not a failure of nerve; it is correct epistemic behavior for an embedded observer facing the dimension it did not and cannot enter of its own will.

Teilhard de Chardin's Omega Point [12]—a convergent attractor toward which the evolution of consciousness tends—has structural resonance with $T(w^*) = w^*$. The present account formalizes this intuition without Teilhard's evolutionary progressivism: the fixed point is a structural condition available to any embedded observer at any moment, not a temporal terminus.

5 Conclusion

The formal structure of *The Imagination Machine* series was arrived at by asking what coherence looks like for an embedded epistemic system. The theological structure described in this paper was arrived at by asking what an ancient formula means when taken literally and what geometry it selects when taken seriously as an epistemic constraint.

They are the same structure.

The hypersphere is the geometry of maximal uncertainty for an embedded observer. The inference–implication loop is the formal expression of what it means to be a cross-section of that structure trying to reflect it accurately. The fixed-point condition is alignment. The irreducibility of will is freedom within a determined geometry. The distinction between generative and compressed inheritance is a philosophy of history in which the development of mathematical language is the slow recovery of inferential machinery from a transmission that began with content it could not yet fully express.

I did not plan this. I noticed it. That is what I have tried to record here.

References

[1] Tracy, M. *The Imagination Machine I: A View from Somewhere.* Unpublished manuscript, Boston University.

[2] Tracy, M. *The Imagination Machine II: Relational Invariants, Quotient Structure, and the Reproducibility of Science.* Unpublished manuscript, Boston University.

[3] Tracy, M. *The Imagination Machine IX: The Moral Principle of Action–Motivation.* Unpublished manuscript, Boston University.

[4] Tracy, M. *The Imagination Machine III: Systems.* Unpublished manuscript, Boston University.

[5] Tracy, M. *The Imagination Machine IV: Linear Encoding of Koopman Spectra in Predictive Agents.* Unpublished manuscript, Boston University.

[6] Tracy, M. *The Imagination Machine V: On Abstraction and Analogy.* Unpublished manuscript, Boston University.

[7] Tracy, M. *The Imagination Machine VI: Holons, Horn Fillings, and the Self-Demonstration of Analogy.* Unpublished manuscript, Boston University.

[8] *Liber XXIV Philosophorum.* Anonymous, c. 12th century. Edited by Françoise Hudry. Turnhout: Brepols, 1997.

[9] Nicholas of Cusa. *De Docta Ignorantia.* 1440. Trans. Jasper Hopkins. Minneapolis: Arthur J. Banning Press, 1981.

[10] Whitehead, A.N. *Process and Reality.* New York: The Free Press, 1929.

[11] Hartshorne, C. *The Divine Relativity.* New Haven: Yale University Press, 1948.

[12] Teilhard de Chardin, P. *The Phenomenon of Man.* Trans. Bernard Wall. New York: Harper and Row, 1959.

[13] Meister Eckhart. *The Complete Mystical Works.* Trans. Maurice O'C Walshe. New York: Crossroad, 2009.

[14] Augustine of Hippo. *De Trinitate.* c. 400–428 CE. Trans. Edmund Hill. Hyde Park: New City Press, 1991.

The Imagination Machine VIII:

The Semiotic Lens and the Justification Loop

Mark Tracy
Boston University
mrktracy@bu.edu

Abstract

This paper develops a phenomenological grounding for the embedded epistemic framework of the Imagination Machine series. Beginning from a simple perceptual question—why do same-sized objects that are farther away appear smaller?—it establishes that the question itself only has meaning through the semiotics of consciousness, and that there is no view from nowhere, no Archimedean point outside of semiosis from which to grasp things in themselves. It then extends this observation to the problem of justification: the standard triad of faith, logic, and experience turns out to be not three alternatives but three aspects of a single recursively self-correcting loop. This loop—in which experience gives meaning to symbols, logic filters incoherent collections of belief, and faith commits to provisional closures that experience then tests—is the phenomenological form of the inference-implication loop. The paper makes no new formal claims. It describes what it is like to be inside the bubble.

1 Introduction

How are propositions ultimately justified? The question has three standard answers: through faith, through logic, or through experience. I want to argue that choosing any one of these options in isolation is misguided—not because the question is meaningless, but because the three cannot be separated. What appears to be a choice between three foundations turns out, on examination, to be a single recursive process in which each term presupposes and enables the others.

To see why, it helps to begin not with epistemology but with perception—with something as simple and immediate as the apparent size of objects at a distance.

2 The Semiotic Lens

Why do same-sized objects that are farther away appear smaller? Is it a consequence of the laws of physics, or is it an accident of my perceptual processing? Or is it impossible for me to know either way?

A naive response is that farther objects appear smaller because the light traveling from an object to the retina forms a cone, and farther objects subtend a lesser angle on the retina.

However, it seems that I could conceivably experience that same raw data of the photons subtending a lesser angle in my retina's receptive field as being objects of equal size but a different clarity, or some other distinction. In other words, an alien perception could operate and subjectively appear entirely differently. So in some sense, this particular appearance of order is an accident of my perceptual processing.

Even more strongly, the question itself only has meaning at all through the semiotics of my consciousness. For the reader who may not be familiar with this language, "semiotics" is the study of symbols and meaning-making. I'll next introduce some basic notions about symbols in order to clarify what I mean by "the semiotics of my consciousness."

A "symbol" as I mean it is composed of two interacting sub-parts: there is the signifier, such as a word (e.g. "arm"), and the signified, which is the actual physical process or processes referred to by the signifier (e.g. my actual arm). The meaning of the signifier—that is, the map from signifier to real, physical processes—is given by a particular brain or mind. What I mean, then, by "the semiotics of my consciousness" is how I lend meaning to symbols, such as "my arm."

Let's dive even further into the example of "my arm." To me, "my arm" refers to everything from the shoulder to the tips of the fingers. First, note that the description I have just given requires knowledge of the meaning of a series of other symbols, such as "shoulder," "fingers," as well as an intuitive understanding of the "from-to" relation. Second, one can easily imagine that to another person, "arm" could mean everything from the shoulder to the wrist, but not including the hand. This example illustrates how every symbol is embedded within a system of symbols, and meaning is assigned to them by a particular mind or brain.

Now, what does it mean that the question of why same-sized objects that are farther away appear smaller "only has meaning at all through the semiotics of my consciousness"? Well, for an object to be "farther" than another object at all refers to a third reference point to which one object is closer and the other farther, and to "appear smaller" necessitates a consciousness at that reference point to whom it appears at all. Finally, and significantly, it assumes the notion of an "object" that is in some sense stable and unified enough to be identifiable as one "thing" across time. So this question only makes sense through the semiotic lens of a consciousness that exists at a point or in a limited region of spacetime upon which stable-enough patterns in physical processes can impinge to impart abstract, object-oriented information.

Though we may now be tempted to say that the apparent fact that same-sized objects that are farther away appear smaller is an accident of perceptual processing, we must also acknowledge that our perceptual processing may itself be a consequence of the "true" laws of physics, or divine Logos, if such exists. For example, it could be that such perception is inevitable, given the reality of some general form of the theory of evolution by natural selection and the survival advantage of such an encoding of information—for example, that it allows us to recognize important spatial information about physical reality.

So then we are led to the conclusion that we cannot possibly know either way whether the apparent fact in question is due to the laws of physics or is an accident of our perceptual processing. The two are irrevocably linked through semiosis, the meaning-making process, itself.

Considering all of this, it seems to me that physics tells us how things go on; but not what goes on. "What goes on" is dependent on consciousness, on abstract, semiotic systems mapping that which in the language of physics may be called "spatiotemporal process" to symbols, or object-oriented, timeless representations.

In other words, to "be something" is to "be-something-to." Physics can tell me how I move my arm, but never what "I" am or what "my arm" is, because those notions are situated within a semiotic system. "My arm" does not physically exist as such. It exists as "arm" only "semiotically," with its meaning mediated by a particular consciousness.

This is of course not to say that physics is not helpful or useful—not at all—but it follows that there is not necessarily ontological privilege for the fundamental "objects" identified by physics. They also can be said to exist as such only semiotically.

Our perceptual experience is always already imbued with meaning—it is not a "raw" or "neutral" input that we then interpret, but comes to us pre-interpreted through learned categories and distinctions. The visual experience of size and distance is one example of this: we don't just passively receive retinal images, but actively construct a meaningful, three-dimensional world of objects located in space.

This meaning-ladenness goes beyond specifically human modes of perception. The broader point is that any organism's *Umwelt* or "lived world" is constituted through its particular ways of making meaning, its semiotic systems. For a bat, the world is primarily a soundscape of ultrasonic reflections; for a dog, a richly textured smellscape; for an electric fish, a field of electrical gradients.

Each organism inhabits a world of significance that is co-constructed through its embodied interactions and evolutionary history. There is no "view from nowhere," no Archimedean point outside of semiosis from which to grasp "things in themselves."

The human case is perhaps special in the degree to which our semiotic systems are flexible, open-ended, and mediated by language and culture. But the basic principle of the semiotic constitution of lived worlds applies across the board. Meaning and being are always entangled: ontology is always bound up with semiosis.

3 Faith, Logic, and Experience

Now the question of justification presents itself with new urgency. If there is no view from nowhere, if perception is always already semiotic, if ontology is always bound up with meaning-making—then how are propositions justified at all? I want to argue that no one of the standard answers provides ultimate justification on its own, and that the attempt to separate them reveals, instead, a single recursive loop.

Faith

By "faith," I mean a commitment or effort to a belief, or a letting go of the question as to whether a particular proposition is justified. It is, in other words, taking a belief as the basis for action in one's life, even without ultimate justification.

To say that everything is ultimately justified through faith may therefore be interpreted as meaning that there is no ultimate justification for any proposition to be found in logic or experience (or elsewhere), and that in the last analysis there is some leap of faith in relying on the truth of any proposition.

Faith as I have defined it here factors into human decision-making in essentially all its forms. In interpersonal relations, we (usually) have faith in the good intentions of our loved ones; in traveling we have faith in the soundness of our infrastructure; and in building technology we have faith in the best scientific understanding of our day.

The case for faith as ultimately necessary finds strong support in the Münchhausen trilemma. The trilemma proposes that the effort to ultimately justify any proposition must terminate with one of three unsatisfactory results: (1) the chain of justification terminates with foundational axioms that are dogmatic and not further justified; or (2) the chain of justification is infinite, with every truth having a prior justification; or (3) the chain of justification closes upon itself, producing logical circularity.

Unless this trilemma can be resolved, it seems that "buying in" to any proposition involves an unavoidable element of faith. And yet, this trilemma reveals that faith isn't strictly speaking a "justification" in itself, but rather a necessary response to the problem of lacking the possibility of ultimate justification through logic and experience alone.

At the same time, faith is a relation between a person and an object of faith. Faith is only "faith-in-something," and the object of faith is only knowable and evaluable at all through experience and logic.

Logic

While the Münchhausen trilemma suggests that there is an indispensable element of faith in believing any proposition, it does not follow that logic has no role in justification at all. Indeed, logic serves a crucial role, along with experience, as part of the framework by which

we decide what to have faith in at all.

Logic is essentially relational. It is not about the absolute truth of propositions themselves, but rather about how the truth value of different propositions relate. Propositions are in turn statements about hypothetical objects' properties and relations.

For example, logic allows us to say that if "all men are mortal," and if "Socrates is a man," then "Socrates is mortal." This is a statement about how the truth value of these statements relate: they all must be true together, or if the conclusion "Socrates is mortal" is false, then it must either be false that "all men are mortal" or that "Socrates is a man." Logic says nothing about whether "all men are mortal" or whether "Socrates is a man"; and as such it says nothing about whether "Socrates is mortal" in any absolute sense.

Indeed, the very meaning of the symbols above are not given in any eternal way. As established in the preceding section, those symbols are ultimately imbued with meaning by an individual consciousness. Objects with properties and relations are abstracted from the fundamentally processual nature of reality. They are unchanging representations that can be said to exist only "semiotically" or symbolically, as a stand-in for a dynamic underlying reality.

In particular, a collection of imaginations in our mind hypothetically maps at a given moment to a single "concept." For example, if I say, "grandmother," you could in principle imagine anything you can possibly imagine and each time say whether it is an instance of the concept "grandmother" or not. But your particular grandmother is or was a dynamic process, not a static, unchanging object.

The relational nature of logic and the role of consciousness in mapping processual reality to symbolic representation are often overlooked but are fairly obvious once one considers them.

The role of logic in justification, then, is to reveal clearly incompatible beliefs. For example, it is logic that allows one to say, "Whatever you mean by 'men', whatever you mean by 'Socrates,' and whatever you mean by 'mortal,' you cannot simultaneously hold that 'all men are mortal; Socrates is a man; and Socrates is immortal.'"

This is a way to make sure that any collection of beliefs is "playing by rules" that constitute it a coherent system of truths, rather than a collection of propositions with no consistent relations between them. Clearly, then, logic plays a key role in creating the "thing" or system of propositions in which one can have faith at all; but it does not provide ultimate justification of the content of propositions.

Experience

It is through experience that symbols are imbued with meaning, and it is experience that gives feedback regarding our faith in systems of propositions. Indeed, faith and logic are themselves experiences or aspects of experience. So then we may be tempted to say that experience

provides the ultimate justification for propositions. However, providing the meaning of propositions and providing feedback regarding the efficacy of faith therein are not the same as "ultimate justification."

To see why, consider how much our experience itself comes to us pre-processed by our concepts, or our "semiotic lens." Our perceptions are already laden with meaning. For example, try to look at these words as symbols without simultaneously seeing the meaning or sound of the word; or even seeing it as a word. It is very difficult to do.

What this implies is that our logic and our other beliefs and assumptions color our experience itself. Experience is not raw data that we receive objectively, according to which we can decide in an absolute sense between propositions. Rather, there is a bootstrapping or recursion between experience giving meaning to symbols and symbols shaping experience itself. This process begins when we are very young, when for example a parent or guardian points to an image of a truck and repeats "truck," and then we see a different vehicle and point to it ourselves and say, "Truck!" It doesn't stop for our whole lives.

At the same time, experience gives us a point of contact with the rest of reality (the "not-me" or the "Other" in my usual parlance). Without experience to provide meaning to systems of logically related symbols, and without experience to give feedback on our commitment to such systems, the whole justification process makes no sense at all.

If experience is theory-laden, then there is some form of faith that colors experience; while logic filters out collections of incoherent propositions in order to exclude unhelpful objects of faith; and experience provides feedback on both our logical arguments and our articles of faith in order to recursively self-correct the whole process.

4 The Loop

What is the ultimate justification for the propositions that I have presented here? I can offer you none. But you can hopefully see that these propositions are logically coherent. Then you can choose to take these ideas on faith. If you do, they can color your experience. For example, once you see the fundamental difference between processual reality and its symbolic representation, or the theory-ladenness of experience itself, you may start to notice differences in your everyday experience. And the results of those changes in experience—perhaps the efficacy of your new beliefs, the way they make you feel, feedback from others—may lead you to reconsider your belief in its logical coherence or your faith in its truth. This may lead you to dialogue: to clarify what was meant by one thing or another, since meaning is ultimately ascribed to these symbols by an individual consciousness.

This process highlights how meaning-making and belief-formation are fundamentally dialogical, pragmatic, and recursively error-correcting without providing ultimate, unassailable justification.

Where does this leave us? What does it mean that justification is a recursively self-

correcting and dialogical process involving provisional faith, logic, and experience, rather than an ultimate "yes or no" status granted by experience alone?

It encourages us to engage actively in this process of belief formation and meaning-making, in the knowledge of what we are doing. It is also to adopt a stance of intellectual humility. This allows us to meet others in a spirit of genuine intellectual curiosity, with the potential for our minds to be changed. It also encourages collaboration and diversity in institutions of meaning-making and belief formation, such as higher education.

The Imagination Machine IX: The Moral Principle of Action–Motivation

Mark Tracy
Boston University
mrktracy@bu.edu

Abstract

We propose an augmentation of Kant's Categorical Imperative in which the object of universalization is not an action alone but a tuple of action and motivation set. The motivation set of an action is the family of minimal subsets of anticipated consequences whose perceived relevance is necessary and sufficient for the action to be chosen. A tuple of action and motivation set is morally admissible if and only if it can be coherently willed to be universally permissible.

1 Explication of Terms

We consider an agent deliberating over actions. The following objects are defined relative to a given decision-making event.

Definition 1 (Action Space)**.** *Let A be the set of possible actions available to the agent.*

Definition 2 (Belief Set)**.** *Let B be the set of equivalence classes of statements of beliefs of the agent, modulo synonymous phrasing. We denote statements using double quotation marks.*

Definition 3 (Relevant Anticipated States of Affairs)**.** *Let C be the set of relevant anticipated states of affairs: those states the agent believes to be made more likely by one possible action than by another. Formally,*

$$c \in C \iff \exists\, a, a' \in A,\ \exists\, b \in B : \quad \text{“}P(c \mid a) > P(c \mid a')\text{”} \in b.$$

The statement “$P(c \mid a) > P(c \mid a')$” reflects the agent's belief. This set captures the states of affairs at issue in the present decision.

Definition 4 (Decision Indicator)**.** *Let $d : A \to \{0, 1\}$ be a one-hot indicator function signaling the action decided upon, so that $d(a) = 1$ if the agent decides to take action a, and $d(a) = 0$ otherwise.*

Definition 5 (Relevance Map)**.** *Let $e : A \to \mathcal{P}(C)$, where $\mathcal{P}$ denotes the power set, associate each action a with the subset of anticipated states of affairs relevant with respect to a:*

$$e(a) = \{\, c \in C \mid \exists\, b \in B,\ \exists\, a' \in A : \text{“}P(c \mid a) \neq P(c \mid a')\text{”} \in b \,\}.$$

Definition 6 (Motivation Set)**.** *Let the* motivation set *M_a of an action a be the family of minimal subsets of $e(a)$ such that, if the agent believed them irrelevant, action a would surely not be chosen:*

$$M_a = \{\, m \subseteq e(a) \mid \exists\, b \in B : \text{“}e(a) \cap m = \emptyset\text{”} \in b \implies d(a) = 0, \\ \text{and} \quad \emptyset \neq m' \subset m \implies m' \notin M_a \,\}.$$

The first condition states that $m \in M_a$ if believing the states in m to be irrelevant would be sufficient to preclude action a. The second condition enforces minimality: no nonempty proper subset of any element of M_a is itself an element of M_a.

Remark 1 (Conjunctive Motivation)**.** *Suppose Carl is choosing between staying at his current job or leaving it to find another, so $A = \{\text{stay}, \text{change}\}$. Suppose that if both a better salary and a shorter commute were believed irrelevant, Carl would surely not change jobs, but if either remains relevant he would be willing to change. Then*

$$\big\{\{\textit{better salary, shorter commute}\}\big\} \subseteq M_{\text{change}}.$$

Remark 2 (Disjunctive Motivation)**.** *Now suppose that if* either *a better salary or a shorter commute were believed irrelevant, Carl would surely not change jobs. Then*

$$\big\{\{\textit{better salary}\}, \{\textit{shorter commute}\}\big\} \subseteq M_{\text{change}}.$$

The minimality condition prevents the redundant inclusion of $\{$better salary, shorter commute$\}$, which would otherwise generate combinatorially explosive supersets.

Definition 7 (Action–Motivation Tuple)**.** *For a given decision-making event, and for the action a for which $d(a) = 1$, the pair (a, M_a) is the* action–motivation tuple*.*

2 The Moral Principle

The Moral Principle of Action–Motivation. Act according to the tuple of action and motivation set which you can simultaneously will to be universally permissible.
No maxim regarding actions alone can be coherently universalized, because one can always contrive a situation in which any action is permissible to prevent a greater evil. The motivation set resolves this by making the object of universalization sensitive to the consequences the agent believes the action to bring about and to the role those anticipated consequences play in the decision. A tuple (a, M_a) is morally admissible if and only if it can be coherently willed that all agents be permitted to perform a whenever their motivation set with respect to a is M_a.

3 Advantages of this Formulation

This formulation allows one to judge the morality of an action both by the nature of the action itself and by what consequences the agent believes the action makes more or less likely. It preserves the formal structure of the Categorical Imperative while resolving its well-known susceptibility to counterexample by actions alone. It is sensitive to the agent's actual deliberative situation rather than to an abstract description of the act.

4 Examples of Universalizable Maxims

The following tuples of action and motivation set are universalizable under the principle:

- Do not lie for the purpose of attaining material personal benefit.
- Do not commit violence for the purpose of attaining material personal benefit.
- Seek out perspectives different from your own for the purpose of better understanding the consequences of your decisions.

- Seek countervailing evidence for the purpose of testing and refining convictions.
- Do not engineer the epistemic closure of others for the purpose of concentrating influence over their world models.
- Do not treat as unwillful what appears willful.

The last two of these are worth dwelling upon.

To coherently will the engineering of others' epistemic closures for the purpose of concentrating influence over their world models (and therefore their actions) would require an agent to license such influence to be exerted upon itself, which would foreclose the very freedom of action that enables the agent to decide in the first place. The loop of universality cannot close.

Will must be accounted for but cannot be exhausted by any formalism, since it is the prerequisite for such a formalism to be created at all. Will must therefore be attributed between minds, not discovered. If to another I can only ever appear willful — that is, if my willing is not externally verifiable to another with absolute certainty — then it would be incoherent to disregard the apparent will of another in my experience, for it would undermine the very position I am at the same time claiming as a willful subject whose deliberation regarding the will of another matters at all.

5 Conclusion

The Imagination Machine series identifies will as the irreducible remainder of the inference–implication loop: the necessity of following and revising a map in territory that no map can fully and faithfully represent. The present paper formalizes the moral condition on that choice. An action–motivation tuple is morally admissible if and only if it can be coherently willed to be universally permissible.

This is not the end of moral discourse but its beginning: an attempt at a minimal condition to foreclose moral hypocrisy, or at least to outline its shape. As a stable world model must survive its own implications when resubmitted to the conditions of its inference, so too must a moral act survive its own implications when resubmitted to the deliberative conditions of its selection. In both cases, the machinery admits internal constraint under the condition of reflexive application.

Self-classificatory reflexivity eventually compels substantive constraint on claims to knowledge and moral action—constraint that is at once limitation and definition.