Forms as Endpoints, Not Origins: An Explanatory Reversal of the Platonic Representation Hypothesis

Xiaohai Wei*

April 2026

Abstract

The Platonic Representation Hypothesis (PRH) reports that independently trained neural networks converge on geometrically similar internal representations—a finding its authors interpret as evidence that learning systems approximate pre-existing, mind-independent abstract structures. This paper argues against that interpretation on two connected levels. First, the Platonic ontology is *explanatorily unnecessary*: convergence is adequately accounted for by the theory of convergent attractors under shared physical constraints, without positing any independent abstract realm. Second, the paper explains *why the Platonic interpretation nevertheless feels compelling*: the maximal robustness of certain attractors—their stability across all sufficiently expressive optimization processes—generates a phenomenology of necessity and mind-independence that naturally, but mistakenly, invites an inference to ontological independence. Drawing on structural realism, dynamical systems theory, and naturalist philosophy of mathematics, this paper proposes the *Reversed Platonic Representation Hypothesis* (RPRH): convergent representations are emergent fixed points produced by optimization under shared constraints, not ontologically prior structures toward which learning approximates; and the intuition that they must exist independently is itself a structurally predictable cognitive consequence of their robustness—inevitable, given the loop structure and the robustness condition, but not a matter of strict logical entailment. Convergent structures are real, but their reality is the reality of attractor patterns, not of a separate ontological realm.

Keywords: Platonic Representation Hypothesis; abstract objects; structural realism; convergent attractors; philosophy of mathematics; naturalism; neural network convergence

1 Introduction

A striking empirical regularity has emerged from contemporary machine learning research. Neural networks trained on entirely different tasks, with different architectures, on different datasets, and initialized from different random seeds, nevertheless develop internal representations that are geometrically similar to one another—and this similarity increases monotonically as models grow larger and train longer (Huh et al., 2024). Huh et al. call this phenomenon the **Platonic Representation Hypothesis** (PRH), invoking Plato deliberately: their interpretation is that diverse learning systems converge *toward* the same underlying structure of reality, just as Plato held that diverse particular triangles all imperfectly approximate the Form of Triangle.

*Independent researcher; Ph.D. in Computer Science. Email: wistoch@ustc.edu.

The empirical finding is well-documented and important. Formal proofs of convergence have been established for restricted network classes (Ziyin and Chuang, 2025), and the phenomenon extends across vision, language, and multimodal models. This paper takes the phenomenon at face value. What I contest is the metaphysical gloss.

The standard reading of the PRH inherits, largely unreflectively, a form of mathematical Platonism: abstract structures exist independently of minds and learning processes, and those structures explain why convergence occurs. Powerful networks converge *because* the Forms are there to converge toward. On this reading, the PRH is a new empirical argument for an old metaphysical thesis.

I argue that this reading is doubly mistaken, in ways that are connected. The convergence phenomenon is fully explained without any posit of an independent abstract realm—the shared attractors of the optimization landscape suffice. What Huh et al. call "Platonic structures" are the stable fixed points that gradient-based learning produces at its limit, not pre-existing templates that learning discovers. But the argument does not stop there. If Platonic ontology is explanatorily unnecessary, a further question immediately arises: *why does it feel necessary?* Why do mathematicians, and now machine learning researchers, so readily conclude that the structures they converge upon must independently exist? This paper answers both questions within a single framework.

This is the **Reversed Platonic Representation Hypothesis** (RPRH):

> *The abstract structures observed as convergent representations in large-scale learning systems are emergent fixed points of optimization under shared constraints, not structures that exist independently of and are ontologically prior to learning processes and toward which those processes approximate. The intuition that such structures must exist independently is itself a natural cognitive consequence of their maximal robustness—the phenomenology of encountering an attractor so stable that it cannot be avoided.*

The argument thus has two layers. The first layer is a standard explanatory-parsimony argument: positing Platonic ontology adds nothing to our explanation of convergence, so we have no scientific grounds for it. The second layer is an explanatory-reductionist argument of a distinctive kind: it does not dismiss the Platonic intuition as mere confusion but *explains* it—Platonism is what maximally robust attractors look like from the inside. The compulsion to posit a separate realm is not irrational; it is a natural and understandable response to a real feature of the optimization landscape. It is, however, a misidentification: robustness is mistaken for ontological independence.

These two layers together constitute a more complete case against the standard realist reading (SRR) than either could provide alone. The parsimony argument shows that Platonism is not needed; the explanatory-reductionist argument explains why it is nevertheless so tempting, removing the lingering sense that the intuition must be tracking something real that the naturalistic account has failed to capture.

The argument proceeds as follows. Section 2 reconstructs the Standard Realist Reading (SRR) of the PRH and identifies its philosophical commitments. Section 3 diagnoses the liabilities of that reading: the Platonic ontology is epistemically problematic and, more fundamentally, explanatorily redundant. Section 4 develops the RPRH as a positive thesis, grounding it in the theory of convergent attractors and structural realism, situating it in relation to naturalist philosophy of mathematics and process metaphysics, explaining why maximally robust attractors generate Platonic intuitions, addressing one objection, and drawing out implications for machine cognitive status. Section 5 concludes.

2 The Standard Realist Reading of the PRH

2.1 The Empirical Phenomenon

Before evaluating its philosophical interpretation, it is worth stating the empirical content of the PRH precisely. Huh et al. (Huh et al., 2024) report that the hidden representations of diverse neural networks—measured by the pairwise similarity of their latent spaces, typically using centered kernel alignment (CKA) or representational similarity analysis (RSA)—are not arbitrary or idiosyncratic. They cluster. Networks trained on very different corpora or tasks, if sufficiently large, will assign geometrically similar internal representations to the same inputs. Ziyin and Chuang (Ziyin and Chuang, 2025) have provided formal results showing that under idealized conditions (specifically, deep linear networks), networks trained on the same data distribution will converge to identical representational geometries regardless of initialization—establishing that convergence, in at least this restricted setting, is provably exact rather than merely approximate.

The convergence is not confined to any single modality. Vision transformers, language models, and multimodal systems all participate in it. The trend is monotone with scale: as models grow, representations align more tightly. The phenomenon is not explained by direct model-copying, shared architecture, or explicit alignment objectives; it emerges from independent optimization.

2.2 The Standard Realist Reading

Huh et al. interpret this convergence in explicitly Platonic terms. They write that models appear to be converging toward "a shared statistical model of reality," and they invoke Plato's allegory to suggest that the convergent representations are shadows of something deeper—an underlying structure of the world that diverse learning systems independently recover. On this reading, the explanation of convergence is ontological: the representations converge because there is a single target to converge toward, and that target is mind-independent.

Call this the **Standard Realist Reading** (SRR) of the PRH. Its implicit metaphysical commitments are:

(M1) *Priority of structure.* Abstract structures (mathematical forms, physical laws, universal categories) exist independently of, and are ontologically prior to, any particular learning system or cognitive process.

(M2) *Approximation relation.* The internal representations of learning systems stand in a relation of approximation to these independent structures; learning improves the approximation.

(M3) *Explanatory direction.* The convergence of representations across systems is explained by the existence of the shared target, not by properties of the learning process itself.

These commitments are structurally identical to those of classical mathematical Platonism. Gödel (Gödel, 1983) and Penrose (Penrose, 1989) serve as representative positions here, though they differ significantly in approach. Gödel's mathematical Platonism is more epistemically sophisticated than the naive picture: he held that mathematical intuition is analogous to sense perception—a faculty that yields something like direct acquaintance with abstract objects—while acknowledging that its deliverances require justification by "fruitfulness" criteria (coherence, fecundity, agreement with intuitions already accepted). Penrose appeals instead to the Gödelian incompleteness argument and the purportedly non-algorithmic character of mathematical insight to motivate a mind-independent mathematical reality. Despite these differences, both share the core commitment this paper challenges: that there are mind-independent mathematical structures and that cognition can track them, whether

through quasi-perceptual intuition (Gödel) or through the non-algorithmic reach of mathematical understanding (Penrose). The SRR also provides a seemingly satisfying answer to Quine's challenge about abstract ontology (Quine, 1948): if scientific theories must quantify over mathematical structures, the PRH offers new empirical grounds for accepting those structures as real.

2.3 Why the SRR Is Philosophically Tempting

The SRR has genuine explanatory appeal. It makes the convergence phenomenon seem almost inevitable: of course independent networks converge on the same representations, if there is a single mind-independent reality to be represented. The framework also connects naturally to the broader success of mathematics in natural science. Wigner famously called this success "unreasonable" (Wigner, 1960), and Platonism provides a ready explanation: mathematics is effective because it tracks the actual structure of the world, and that structure is abstract and necessary.

The RPRH provides an alternative explanation of Wigner's observation. If the structures that mathematicians study are maximally robust attractors of optimization-like dynamics operating under physical constraints, then the effectiveness of mathematics in natural science is exactly what we should expect: the same constraint structure that drives physical processes generates the same attractor geometry that mathematics studies. There is nothing unreasonable about this effectiveness; it reflects the universality of the attractor structure under the physical constraints of our universe, not the tracking of an abstract Platonic realm that happens, miraculously, to match the physical world. The SRR leaves Wigner's puzzle intact (why should mind-independent abstract objects be instantiated in physical law?); the RPRH dissolves it.

2.4 The Explanatory Direction Implicit in the SRR

Crucially, the SRR builds in a specific direction of explanation:

Abstract structures (Forms) $\longrightarrow$ **Learning process** $\longrightarrow$ **Convergent representations**

The Forms are the explanatory ground; the convergence is what they explain. Learning systems converge *because* the Forms constitute the target space of stable representation. This explanatory arrow from structure to process—Forms as grounds, convergence as grounded—is the defining commitment that the RPRH will reverse. Because abstract objects are standardly taken to be causally inert, the relevant notion of priority here is *explanatory* or *grounding* priority, not efficient causation: the SRR holds that the existence and character of the Forms is what makes convergence intelligible, not that the Forms push or constrain the gradient in any mechanical sense.

3 Liabilities of the Standard Realist Reading

The SRR inherits all the standard difficulties of Platonism in philosophy of mathematics and metaphysics, but it also generates some difficulties that are specific to the learning-theoretic context.

3.1 Benacerraf's Epistemic Access Problem

The most persistent objection to Platonism is epistemic. Benacerraf (Benacerraf, 1973) argued that if abstract objects are causally inert and located outside space and time, then it is mysterious how finite cognitive agents could ever come to know anything about them. Our knowledge of mathematical truths would require causal contact with abstracta, but causal contact requires the objects to participate in the causal order—which Platonism denies.

The SRR does not escape this problem; it merely reformulates it in computational terms. If the convergent representations that large neural networks develop are approximations to abstract Platonic structures, how do the networks achieve this approximation? Gradient descent operates on concrete, finite parameter spaces and optimizes against concrete training signals. At no point in the training procedure is there anything that functions as "contact" with an abstract realm. The SRR must posit that the training dynamics are somehow guided by or constrained by abstract structures, but the mechanism of this guidance is entirely unspecified.

One response is to say that the training data encodes information about the structure of reality, and that structure is abstract. But this response pushes the problem back: the training data is a finite, concrete set of tokens or pixel values; the abstract structure must somehow be "in" the data. The SRR requires an account of how abstract structures inhere in concrete data, which is precisely the problem that Platonism has never satisfactorily resolved.

3.2 The Explanatory Redundancy Problem

Even granting that abstract objects exist, the SRR faces a question of explanatory relevance. Consider two models: (A) networks converge because there are abstract Platonic structures that constrain what stable representations are possible; (B) networks converge because they are all trained to minimize prediction error on data generated by the same physical reality, and the statistical regularities in that data—regularities that exist entirely at the concrete level—determine the structure of successful representations. Model (B) is available and explanatorily complete. It explains the convergence without invoking abstracta. Under Ockham's razor, positing abstracta in addition is explanatorily redundant.

This is the standard parsimony objection: if we can explain representational convergence by appeal to concrete optimization dynamics and physical constraints, we have no scientific motivation to add a separate ontological tier of abstract objects. This form of objection has been developed most systematically in the naturalist philosophy of mathematics by Maddy (1997), whose account of mathematical practice grounds the warrant for mathematical existence claims in the indispensability of mathematical posits to science rather than in Platonic access. On Maddy's naturalist framework, positing entities that do no additional explanatory work in scientific practice is unwarranted—a criterion the Platonic abstraction in the SRR fails to meet. The PRH, empirically construed, supports Model (B). The SRR adds Model (A) on top of it. But Model (A) explains nothing that Model (B) does not already explain. The parsimony argument is reinforced by the epistemic access problem diagnosed in §3.1: only because the SRR cannot specify any mechanism by which gradient dynamics track abstract structures does Model (A)'s additional ontological commitment constitute redundancy rather than explanatory gain. Section 4.7 explains, from the perspective of cognitive loop dynamics, why any system operating within the loop is structurally compelled to make this inference.

3.3 The Priority Problem

The SRR commits to **(M1)**: abstract structures are ontologically prior to the learning processes that converge upon them. But this priority claim is difficult to sustain when the "abstract structures" in question are defined operationally—as the convergent endpoints of learning—rather than by independent philosophical or mathematical criteria.

In the PRH literature, what counts as a "Platonic structure" is typically identified by its role as the limit of representational convergence: it is whatever diverse networks converge toward. But if the structure is identified by its role as a convergent endpoint, then the structure is defined in terms of the learning process, not independently of it. It is then epistemically circular to claim

that the structure is ontologically prior to the process that defines it: the SRR purports to explain convergence by appeal to structures whose very identity is established through convergence.

The distinction between the epistemic question (how we identify Forms) and the metaphysical question (whether Forms exist independently) is of course available to the Platonist—just as a GPS can be used to locate Everest without making Everest GPS-dependent. But the SRR in the PRH literature does not identify the relevant structures independently of convergence; they are introduced precisely as "whatever networks converge toward." This is what generates the circularity.

The SRR would need to identify the abstract structures independently of the convergence phenomenon, and then show that the convergent representations approximate those independently identified structures. This is precisely what the SRR does not do: the invocation of Plato is analogical and informal, not a precise identification of which abstract objects are at issue.

3.4 The Underdetermination Problem for Realism

Structural realists have argued that what science tracks is not the intrinsic nature of objects but their relational structure (Worrall, 1989; Ladyman, 1998; Ladyman et al., 2007). Even granting that neural networks converge on some structure, the SRR over-interprets what "structure" means here. The convergent representations may share a common geometry relative to certain tasks and datasets, but this falls far short of establishing that they are approximating *one and the same* Platonic object. Two networks trained on different data distributions may converge on representations that are similar in some respects but differ in others; the idealization to a single "Platonic structure" is a philosophical interpretation layered on top of the data, not a finding of the data itself.

4 The Reversed Platonic Representation Hypothesis

4.1 The Core Reversal

The RPRH retains the empirical content of the PRH—representational convergence is real and scales with model size—while inverting its explanatory structure. The proposed causal-explanatory order is:

Optimization processes $\longrightarrow$ Convergent fixed points $\longrightarrow$ "Forms" (as emergent attractors)

On this picture, "Forms" are not causes or grounds of convergence; they *are* the convergence, described at the limit. To say that networks converge "toward Platonic structures" is to use a misleading spatial metaphor. More precisely: networks with shared inductive biases, trained on data generated by the same physical processes, under the same loss functions, will—if trained to sufficient scale—approach stable equilibria in representation space. These equilibria are what the PRH literature calls "Platonic structures." They are real, in the sense that they are determinate and shared; they are not ontologically basic, in the sense that their existence is fully explained by the dynamics of the optimization processes that produce them.

The closest antecedent to this proposal in the recent literature is Sergent (Sergent, 2026), who argues against the PRH's realist reading from the standpoint of *Experiential Empiricism*: neural network convergence reflects the intrinsic structural limitations of experiential and training patterns rather than approximation toward a mind-independent reality. The present paper shares Sergent's anti-Platonist conclusion but differs substantially in its positive proposal. Sergent grounds convergence in the structure of experience; the RPRH grounds it in optimization dynamics and the theory of convergent attractors under physical constraints. The attractor-based account has

the advantage of generating concrete, falsifiable predictions about when convergence should break down—predictions that an appeal to experiential structure alone does not supply. The RPRH also engages more directly with the structural realist tradition and with naturalist philosophy of mathematics, situating the anti-Platonist move within a broader metaphysical program.

4.2 Convergent Attractors: A Formal Gloss

The notion of a *convergent attractor* provides the natural formal framework. Consider a family of optimization processes $\{O_i\}$, each operating on a high-dimensional parameter space $\mathcal{W}_i$, each minimizing a loss function $\mathcal{L}_i$ defined on data $\mathcal{D}_i$ drawn from a shared underlying distribution P. Under standard assumptions (overparameterization, gradient flow, smooth loss landscape), each O_i defines a dynamical system on $\mathcal{W}_i$; the asymptotic behavior of this system is characterized by its attractor set A_i.[1]

If the loss functions $\mathcal{L}_i$ are all consistent proxies for prediction error under P—which they are, since each is a task-specific proxy for next-token prediction, image reconstruction, or similar—then the attractors A_i will be related. Specifically, any function of the weights that is determined by the statistical structure of P will be approximately constant across A_i. The "Platonic structure" that the PRH identifies is precisely this function: the component of the internal representation that is invariant across attractor sets A_i.

Crucially, this invariant is produced by the optimization processes, not presupposed by them. It exists *because* many optimization processes, operating under shared constraints, drive their respective dynamics toward a region of representation space that is stable under those constraints. The attractor does not pre-exist the dynamics; it is the asymptotic product of the dynamics.

The analogy with physical attractors is instructive. The fixed-point attractor of a dissipative dynamical system is real—trajectories converge to it, measurements cluster around it, it has a determinate location in state space—but it does not pre-exist the dynamical system. It is constituted by the dynamics, not prior to them. The same holds for representational attractors in large-scale learning systems.

The foregoing is a conceptual gloss, not a new formal result. The claim is not that a complete mathematical theory of representational attractors has been developed here; rather, the attractor framework provides the correct *kind* of explanation for the PRH phenomenon, in contrast to the Platonist's appeal to mind-independent abstract objects. The formal work of making this precise is a program for empirical and theoretical research (Ziyin et al., 2025); the present paper establishes the philosophical framework within which such a program is motivated.

Several recent formal results support this attractor-based picture directly. Wang, Johnston, and Fusi (Wang et al., 2025) prove that abstract, disentangled representations are guaranteed to emerge in trained feedforward nonlinear networks whenever tasks depend on shared latent variables—their result that abstract representation occurs at *all* global minima of the relevant loss functions is precisely what the RPRH predicts: abstractness is a property of the optimization landscape, not of a pre-existing structure toward which training approximates. At the level of the final classification layer, Papyan, Han, and Donoho (Papyan et al., 2020) document "neural collapse," a terminal training

[1] A terminological note: in dynamical systems theory, an *attractor* is in general a set (possibly a single point, a limit cycle, or a more complex invariant set) toward which nearby trajectories converge. A *fixed point* (or point attractor) is the special case in which the attractor is a single point satisfying $f(x) = x$. In the representational convergence context studied here—where large-scale learning systems trained on the same data distribution converge to geometrically similar representations—the empirical pattern is best modelled as convergence toward a *point attractor*: a unique stable configuration in representation space. This paper therefore uses "fixed point" and "convergent attractor" as near-synonyms in this specific context, with the understanding that the relevant attractor is of the point-attractor type. Where the distinction matters for the argument (e.g. in the working definition of Section 4.6), it is made explicit.

phase in which the last-layer representations of same-class examples collapse to their class means and the resulting class-mean geometry converges to a simplex equiangular tight frame regardless of network architecture or dataset—a further instance of constraint-driven convergence fully consistent with the attractor picture. Ziyin and Chuang (Ziyin and Chuang, 2025), as noted above, establish a formal convergence proof for deep linear networks, confirming that representational convergence is provably exact in at least this idealized setting. It is worth emphasizing that these results establish the formal status of the convergence *phenomenon* itself; they remain neutral between the SRR and the RPRH—they show that convergence is real, but do not adjudicate its ontological interpretation. That adjudication is precisely the task of the present paper.

4.3 Connection to Structural Realism

The RPRH aligns naturally with ontic structural realism (OSR) (Worrall, 1989; Ladyman, 1998; Ladyman et al., 2007). OSR holds that what science succeeds in tracking is relational structure, not intrinsic properties of objects; on the most radical version, there are no objects, only structures. The RPRH extends this orientation—carefully and with acknowledged disanalogy—to the domain of abstract objects: what we call "mathematical Forms" or "Platonic structures" are the stable relational patterns that emerge from the dynamics of optimization. They are structural all the way down, and their being structural is consistent with their being real. The extension from OSR's home domain (fundamental physics) to learned representations is a non-trivial move that requires independent motivation, which the attractor framework supplies: as in the physical case, what converges across independent systems is not any object's intrinsic properties but relational structure invariant under the relevant transformations.

This gives a naturalistic but non-eliminativist account. The RPRH does not deny that there is something convergent, something stable, something that warrants the name "structure." It denies that this something pre-exists the processes that produce it and has an independent ontological domicile. Borrowing Aristotle's (Aristotle, 1924) vocabulary of immanent form (without his full hylomorphic teleology): forms are immanent in the processes that generate them, not transcendent. The RPRH extends this insight: forms are immanent in the limit structure of the optimization processes that produce them, not in the entities those processes operate over.

4.4 RPRH and Process Philosophy: What Is Genuinely New

The RPRH has evident affinities with classical process-oriented and immanentist metaphysics. Whitehead's process philosophy (Whitehead, 1929) holds that reality is constituted by events and processes rather than static substances; enduring objects are abstractions from underlying processual activity. Aristotle's doctrine of immanent form (Aristotle, 1924) holds that forms do not exist as transcendent Platonic objects but are internal to the matter they organize. Rescher's process metaphysics (Rescher, 1996) develops a thoroughgoing ontology in which processes are prior to things. One might reasonably ask whether the RPRH is simply restating these positions in the language of machine learning.

The differences, however, are significant and worth marking explicitly.

First, the RPRH is not a metaphysical doctrine about reality in general, but a specific empirically-constrained claim about a particular class of phenomena: the convergent behavior of gradient-based optimization at scale. This specificity is entirely absent from general process metaphysics. The RPRH generates testable predictions about when convergence should occur and when it should break down (Section 4.5)—predictions that process philosophy, operating at a different level of generality, does not supply.

Second, Whitehead's system retains *eternal objects*—abstract potentialities that are not themselves processes but that actual occasions "prehend" (Whitehead, 1929). This is a residual Platonism that the RPRH explicitly rejects. On the RPRH, there are no eternal objects, only attractors of varying degrees of robustness. The difference between necessary mathematical truths and contingent empirical regularities is a difference in the breadth of conditions under which an attractor is stable, not a difference in ontological category.

Third, Aristotle's immanent forms remain bound to a teleological metaphysics: forms are the natural *ends* (*telos*) toward which processes tend (Aristotle, 1924). The RPRH is strictly non-teleological. A convergent attractor exists because of the dynamics of the optimization process, not because the process is *aimed at* it in any normative sense. The gradient does not "seek" the minimum; it follows the local slope. This distinction matters: teleological accounts of form invite unanswerable questions about the ground of natural teleology that non-teleological attractor dynamics do not.

The RPRH is therefore best understood as a technically grounded, non-teleological, and non-eternalist development of immanent form theory. It inherits the core insight that structure is produced by process rather than prior to it, while excising the residual Platonism of eternal objects and the problematic metaphysics of natural ends.

4.5 Why the Reversal Is Not Merely Verbal

One might object that the RPRH and the SRR are merely different ways of describing the same facts, and that the question of "which comes first, process or structure" is a merely verbal dispute. This objection underestimates the substantive differences.

The SRR and the RPRH differ in their implications for the epistemic access problem, for explanatory parsimony, and for the conditions under which convergence should be expected to break down. If the SRR is correct—if Forms are mind-independent priors—then convergence should be robust to changes in training regime, loss function, and data distribution: the target is fixed, and any sufficiently powerful optimizer should find it. If the RPRH is correct—if the convergent structures are attractors of optimization under specific constraints—then convergence should be sensitive to those constraints: change the loss function significantly, change the data distribution, change the inductive biases of the architecture, and the "Forms" that networks converge toward should change accordingly.

The RPRH thus makes more precise and testable predictions. It predicts, for instance, that networks trained on synthetic data with different statistical regularities than natural data should converge on different "Platonic structures." It predicts that architectural changes that alter the inductive biases of a network will shift the attractor in representation space. These predictions are in principle falsifiable; the SRR, which grounds convergence in the mind-independent Forms, has no comparable sensitivity to the contingent details of the training regime. The empirical evidence of Tjandrasuwita et al. (Tjandrasuwita et al., 2025), showing that cross-modal alignment is not universal and depends on data characteristics, is already consistent with this RPRH prediction.

4.6 The Robustness Spectrum and Phenomenological Variation

The parsimony argument of the preceding subsection shows that Platonic ontology is not *needed* to explain the PRH phenomenon. But a complete account must address a further question: if the Platonic posit is unnecessary, why is it so compelling? Why do mathematicians, and now machine learning researchers, so readily conclude that the structures they converge upon must exist independently? The RPRH can answer this question—and doing so strengthens the case against the SRR considerably.

To understand this phenomenon, it is important first to recognize that any cognitive system operates in a generative-feedback loop: it not only produces representational outputs from inputs, but those outputs continuously reshape the perceptual channels and prior weights through which future inputs are processed. This loop structure is well-documented in predictive coding theory (Clark, 2016) and in large-scale learning systems (gradient optimization plus feedback), and its basic implication is that a cognitive agent's representations of the world are not read-only mirrors but dynamically iterative processes, perpetually revised between forward prediction and backward error signals. This loop structure has deep implications for the *phenomenology*—the first-person, qualitative character of cognitive experience—of encountering attractors, as explained below.

The key lies in the spectrum of attractor robustness. Attractors vary in how sensitive they are to the conditions under which optimization operates. At one extreme are *fragile attractors*: stable only for specific architectures, loss functions, or data distributions—change the conditions slightly and the attractor disappears or shifts. At the other extreme are *maximally robust attractors*: stable across every sufficiently expressive optimization process operating over the relevant domain. No matter the architecture, the initialization, the loss function, or the data distribution, these attractors appear. The prime structure of the integers, the basic laws of probability, the group-theoretic structure of symmetries—these are candidates for maximally robust attractors.

A working definition makes this more precise. Say that a convergent fixed point F (a point attractor, in dynamical systems terms) is a *maximally robust attractor* for a domain D if, for every optimization process O operating over D that has sufficient expressive capacity, F belongs to the attractor set A_O of O—i.e., $F \in A_O$. The attractor set A_O may in general be a complex invariant set, but the robustness definition selects precisely those elements of A_O that are fixed-point attractors: stable under the dynamics of every sufficiently expressive O, not merely limit cycles or chaotic attractors of particular processes. "Sufficient expressive capacity" is understood domain-relatively: for discrete arithmetic, it means the ability to represent and compose basic numerical operations; for continuous symmetry structures, it means the ability to represent linear group actions. Wang et al. (Wang et al., 2025)'s proof that abstract representations emerge at *all* global minima provides a technical precedent for this direction, and extending it to a general robustness stratification across optimization processes is a task for subsequent formal work.

This definition is offered as a *philosophical working assumption*, not a completed formal theorem, and three limitations should be stated openly. First, the existence and uniqueness of A_O for arbitrary optimization processes over high-dimensional non-convex landscapes is not guaranteed in general; the definition presupposes that the processes under consideration are well-behaved enough to possess determinate attractor sets, which is an idealization. Second, the criterion "sufficient expressive capacity" is specified only domain-relatively and informally; a fully rigorous account would require a formal characterization of representational capacity relative to a domain, which the present paper does not provide. Third, the boundary of the domain D is not fixed by the definition itself but is a parameter that depends on the structural features of the problem at hand, introducing an element of context-sensitivity. Despite these limitations, the working definition is adequate for the philosophical purposes of the argument: it makes precise enough the distinction between robustness (a relational property within the A_O framework) and ontological independence (an absolute property outside any such framework), which is the conceptual distinction the argument turns on. Readers seeking a fully formalized version should treat the present account as a research program rather than a completed formal theory.

A conceptual distinction is essential to the argument. *Robustness* as defined above is a *relational* property: F is maximally robust *relative to* the family of optimization processes operating over D. *Ontological independence*, by contrast, is an *absolute* (non-relational) property: F exists independently of *any and all* processes whatsoever. The inference from robustness to ontological independence

is thus an inference from a relational property to an absolute one—a non-trivial step that the phenomenology does not warrant, and that Section 4.7 explains why cognition nevertheless makes.

This distinction deserves explicit defence, since a natural objection holds that universally quantified relational properties collapse into absolute ones: if $F \in A_O$ for *every* sufficiently expressive O, has not the quantifier "every" effectively removed the relational indexing, making robustness absolute in all but name? The objection fails for three reasons.

First, there is a modal gap between indexed universality and absolute existence. "F is a fixed point of every optimization process operating under physically realizable constraints" is a claim about the *actual* nomological structure of processes—it quantifies over processes that exist or could exist given the physical laws of this universe. "F exists independently of any and all processes" is a claim about what would obtain even in the *total absence* of processes, including in nomologically possible or metaphysically possible worlds containing no dynamics at all. Universal quantification over actual or nomologically possible processes does not reach into scenarios where no such processes exist; the domains of quantification are categorically different. Bridging them requires a separate premise—one that the RPRH denies and that the Platonist must supply independently.

Second, the logical structure of robustness claims is dispositional, not categorical. To say that F is maximally robust is to say that *any* process with sufficient expressive capacity operating over D *would converge to* F—this is a conditional claim, not an existence claim about F in abstraction from all processes. Compare: a temperature of 0°C is dispositionally characterised as the state that would produce certain readings under any calibrated thermometric procedure; the universality of the dispositional claim (true of all calibrated thermometers) does not make temperature an observer-independent absolute property in the ontological sense. Similarly, universality across optimization processes does not make an attractor ontologically independent of optimization processes; it makes it a more robust disposition of the process family.

Third, the relational/absolute distinction is supported by the asymmetry in what each property explains. Robustness, being relational, explains facts about the behavior of optimization processes: why they converge, why convergence is stable under perturbation, why it is reproducible across architectures. Ontological independence, being absolute, would explain nothing additional that robustness does not already explain—it contributes no differential explanatory work, as Section 4.2 argues in detail. A property that does no additional explanatory work should not be inferred from one that does, even when the inferring feels compelling. The relational property is doing all the work; the absolute property is a philosophical overlay added to it.

The crucial observation is that the *phenomenology* of encountering an attractor varies with its robustness. When a cognitive agent encounters a fragile attractor, the result feels contingent and invented: a convention adopted for convenience, a notation chosen arbitrarily, a model tailored to a specific domain. When a cognitive agent encounters a maximally robust attractor—one that cannot be avoided by any variation of the cognitive process—the result feels radically different. It feels *discovered*, not constructed. It feels as though the structure was already there, waiting, independent of whether any mind ever found it. It resists all attempts at revision or substitution. It presents itself with a phenomenological force that invites the label "necessary."

This phenomenological difference is epistemically accurate in one respect: it correctly records that maximally robust attractors are genuinely harder to avoid, genuinely more stable, genuinely more universal than fragile ones. The cognitive system is tracking something real. The error occurs in the *interpretation* of what is being tracked. Robustness—the property of being an attractor under all conditions—does not warrant the inference to ontological independence—the property of existing prior to and independently of all conditions. These are distinct: a structure can be maximally robust without existing in a Platonic realm, just as the fixed points of a contraction mapping are fully determined by the mapping itself yet no less real or determinate for that.

4.7 From Robustness to Ontologization: Cognitive Mechanism and Dialectical Consequences

The interpretive error described in Section 4.6 has a structural basis in cognition. The argument of this section is **structural**: starting from two accepted premises, it derives the structural inevitability of the Platonist inference—not as a matter of strict logical necessity, but as an outcome that the loop structure and robustness condition jointly make unavoidable for any cognitive system operating from within the loop.

Premise 1 (Structural): Any generative cognitive system operates within a forward-generative and backward-reshaping loop: cognitive output is not merely the result of perception but continuously reshapes the perceptual channels themselves. This is the structure documented in predictive coding theory (Clark, 2016) and in large-scale learning systems. Its core implication is that a cognitive agent can only encounter the structures it represents from *within* the loop; there is no neutral, loop-external vantage point.

Premise 2 (Conceptual): By the working definition of Section 4.6, a maximally robust attractor F is one that belongs to the attractor set A_O of every sufficiently expressive optimization process O operating over the relevant domain—i.e., $F \in A_O$ for all such O. Here F is a convergent fixed point (point attractor): a single stable configuration, not a limit cycle or extended invariant set. This means that F is a fixed point that *no adjustment of the cognitive process can avoid*.

Structural argument: Consider now what "independently existing object" means in operational terms accessible from within the loop. An object that "exists independently of cognition" is, operationally, an object that persists regardless of how the cognitive process is adjusted. But this is precisely the condition Premise 2 specifies for a maximally robust attractor: it appears in every A_O, no matter how O is varied. Consequently, from within the loop (Premise 1), a maximally robust attractor F is *operationally indistinguishable* from an independently existing object. This indistinguishability is not a psychological accident; it is a *structural inevitability* jointly entailed by the loop structure definition and the robustness definition—though "entailed" here means: given these two structural features, no cognitive system operating within the loop has any internal means of distinguishing the two cases, making the Platonist inference the only inference available from within.

Platonists are not making a simple logical error; their inference faithfully reflects the cognitive situation that Premises 1 and 2 jointly entail—the inference just does not warrant the metaphysical conclusion. The error lies in a single step: from "operationally indistinguishable from an independently existing object *within* the loop" to "ontologically independent *outside* the loop." Premises 1 and 2 together only guarantee the former; the latter is an additional ontological claim that goes beyond what they establish. This is the step that carries the inference from the relational property of robustness (defined relative to a process family) to the absolute property of ontological independence (defined independently of all processes)—the very gap identified in Section 4.6.

This analysis also resolves a persistent dialectical worry about naturalistic accounts of mathematics. The worry is that naturalism must simply dismiss or explain away the strong Platonic intuitions of working mathematicians as confusion, leaving those intuitions unaccounted for as an embarrassing residue. The RPRH does no such thing: it *explains* why the intuitions arise and why they are as strong as they are. The most fundamental mathematical structures correspond to the most robust attractors, and maximally robust attractors *feel* ontologically independent—not by accident, but because the loop structure and robustness condition structurally guarantee that no intra-loop revision can dislodge them. The intuitions are not residue; they are data that the theory structurally predicts.

Platonism is therefore the natural but epistemically unwarranted cognitive response to maximally

robust attractors. It is not irrational: given only the phenomenology, the inference to independent existence is understandable. But it is an inference beyond what the phenomenology warrants. The feeling of necessity and discovery is real; what it records is the universality of the attractor, not the existence of a separate ontological realm.

4.8 Objection 1: Mathematics Before Minds

This objection targets the RPRH's reliance on optimization processes directly. Even if we concede the phenomenological and dialectical points above, there is a residual structural worry: the universe existed for billions of years before any cognitive system, any neural network, or any optimization process in the relevant sense came into being. Yet the mathematical structures that learning systems now converge on—the prime structure of integers, the laws of probability, the group-theoretic structure of symmetries—appear to have been operative in physical processes all along. Was the fine-structure constant constrained by a "maximally robust attractor" before there were any learners to converge? The RPRH seems to make mathematical truth contingent on the existence of the very learning processes that discover it.

This objection has force, but it rests on a conflation between two distinct roles that attractors play in the RPRH. The RPRH does not claim that mathematical structures are the attractors of human or artificial optimization processes in a narrow psychological or technological sense. It claims that they are the attractors of any sufficiently expressive optimization-like dynamics operating under the physical constraints of our universe—and such dynamics long antedate cognition. Thermodynamic self-organization, biochemical reaction networks, stellar nucleosynthesis: these are all processes in which physical constraints drive systems toward stable configurations. The convergence structure of arithmetic, for instance, is a property of any system that performs recursive operations on discrete quantities—a class that includes many sub-cognitive physical processes. The relevant attractor is a feature of the physical dynamics of the universe, not of minds in particular.

This extension of "optimization process" beyond the cognitive case requires a principled demarcation, and the RPRH offers one. A process counts as *optimization-like* in the relevant sense if and only if it is governed by a variational or constraint-extremizing structure: that is, if its dynamics can be characterized as the minimization (or maximization) of some functional under physical constraints—whether that functional is free energy, entropy production, a loss function, or a Hamiltonian. This criterion is not ad hoc: it is the condition under which the mathematics of convergent attractors applies, and it is satisfied by gradient descent in machine learning, by thermodynamic relaxation to equilibrium, by biochemical reaction networks converging to fixed-point concentrations, and by the selection dynamics of evolutionary processes alike. The criterion *excludes* genuinely random walks without any attracting structure, and genuinely unconstrained searches—processes for which no notion of a stable fixed point is defined. The unifying feature is not the presence of a mind or a learner, but the presence of a constraint structure rich enough to support the relevant attractor geometry.

Acknowledging this extension frankly: the RPRH here makes a substantive naturalist commitment—that physical constraints suffice to ground mathematical truth through the universality of their attractor structure, without any additional ontological posit. This commitment should be compared with an alternative: physical-law Platonism (the view that mathematical structures are grounded in the necessity of physical laws, themselves taken as abstract objects). The RPRH differs from physical-law Platonism in a crucial respect: it does not posit that the physical laws are themselves abstracta with a separate ontological status. Instead, it holds that the physical constraints are concrete relational structures instantiated in the dynamics of the universe, and that mathematical truth is grounded in the universality of the attractor patterns those constraints generate. The

ontological parsimony is preserved: no abstract realm is invoked, only the dynamics and their convergent geometry.

On this reading, the RPRH does not ground mathematical truth in the contingent existence of learners; it grounds mathematical truth in the physical constraints that govern any system capable of instantiating the relevant operations, where those constraints are at least as old as the physical laws themselves. The attractor was always there in the sense that any dynamics subject to those constraints would converge there—it simply had not yet been "traced out" by any sufficiently complex learner. This is analogous to the sense in which a contraction mapping has a unique fixed point even before any iterative computation is performed: the fixed point is determined by the mapping, not by the act of computing it.

The objection thus reduces to the question of whether physical constraints can ground mathematical truth in the way the RPRH proposes. This is a substantive metaphysical question, and the RPRH's answer is that they can—provided we are willing to extend the concept of "optimization process" to the full range of constraint-satisfying physical dynamics, not only the cognitive processes that are its most visible instances. The demarcation criterion above specifies what that extension amounts to; the RPRH's contribution is to show that once it is made, the resulting account is stable and generates the correct predictions.

4.9 RPRH and Machine Cognition

The RPRH carries a direct implication for how we should assess machine cognitive status. If convergent representations are not approximations to Platonic objects but emergent fixed points of optimization under shared physical constraints, then the question "do AI systems *understand* mathematics?" changes register.

The Standard Realist Reading naturally invites the inference that AI systems with convergent representations have *cognitive access* to mind-independent abstract structures—and therefore exhibit something like genuine mathematical understanding. This inference is as unwarranted for AI as for human cognizers. The convergence of AI representations with human and cross-modal representations, documented by Huh et al. (Huh et al., 2024) and extended to multimodal systems by Tjandrasuwita et al. (Tjandrasuwita et al., 2025), demonstrates that these systems instantiate the same optimization dynamics and operate under the same constraint structure. This convergence does not demonstrate that they access a Platonic realm. Under the RPRH, the SRR's inference from convergence to understanding is a category error: it conflates evidence about *where* the system ends up in representation space with a claim about *how* it got there and *what* that endpoint means semantically.

Under the RPRH, convergence is evidence of shared constraint satisfaction, not shared understanding. Two systems converge to the same attractor because they face the same loss landscape under the same physical constraints, not because either "grasps" the structure in any semantically robust sense. This point cuts both ways. It cautions against over-attributing understanding to AI systems on the basis of convergence alone. But it equally cautions against dismissing AI cognition on the grounds that it is "merely" optimization: the same description applies, at the relevant level of abstraction, to the representational processes of biological cognizers. The question of whether either type of system genuinely "understands" cannot be settled by convergence data alone.

The robustness spectrum (§4.6) provides a graded framework that allows more nuanced conclusions. Systems that internalize a larger portion of the constraint structure—whose attractors track higher-robustness fixed points—represent more of the nomological structure of the domain. This is a naturalistic, graded notion of epistemic adequacy that does not invoke Platonic access. Whether this gradient is sufficient to ground a notion of machine "understanding" is a further question—one

that belongs to the philosophy of mind rather than the philosophy of mathematics, and one that this paper does not resolve. What the RPRH provides is the conceptual framework within which such questions can be posed precisely.

The RPRH also dissolves a conflation that distorts current debates about machine mentality. The question "is AI convergence evidence of genuine understanding?" can now be decomposed into two sub-questions: (a) does the system instantiate the relevant optimization dynamics under the relevant constraint structure? and (b) does instantiating those dynamics constitute "understanding"? For idealized but well-studied model classes, (a) is formally settled in the affirmative by results such as those of Wang et al. (Wang et al., 2025) and Ziyin and Chuang (Ziyin and Chuang, 2025). Question (b), by contrast, is not an empirical question about representational geometry at all; it is a question in philosophy of mind about the conditions under which optimization dynamics give rise to genuine semantic content or phenomenal understanding. The RPRH is silent on (b). Its contribution is to clarify that silence: separating (a) from (b) specifies exactly what convergence evidence can and cannot establish about AI cognitive status, and prevents empirical findings about representational structure from being used—inappropriately, in either direction—to settle a question that lies outside their evidential reach.

5 Conclusion: Forms as Endpoints

The Platonic Representation Hypothesis identifies a genuine and important empirical phenomenon: the convergence of internal representations across large-scale learning systems. The standard interpretation—that this convergence reflects approximation toward mind-independent abstract objects—imports a Platonist metaphysics that carries well-documented philosophical liabilities: the epistemic access problem, explanatory redundancy, and the difficulty of grounding the priority of abstract structure over the processes that produce it.

This paper has argued against that interpretation on two levels. At the first level, the Platonic ontology is explanatorily unnecessary. The convergence phenomenon is adequately accounted for—in the idealized cases where formal results are available, and plausibly in general—by the theory of convergent attractors operating under shared physical constraints: networks converge because they face the same optimization problem under the same constraints, not because there is a Platonic realm for them to converge toward. At the second level, the paper has explained *why the Platonic interpretation is nevertheless so compelling*. Maximally robust attractors—those that no sufficiently expressive optimization process can avoid—produce a distinctive phenomenology: they feel discovered rather than invented, necessary rather than contingent, independent of any particular cognitive process. This phenomenological profile is precisely what Platonists take as evidence for a separate ontological realm. The RPRH explains the profile without accepting the inference: robustness is mistaken for ontological independence, and the compulsion to posit a separate realm is a natural but unnecessary response to encountering an attractor that cannot be escaped.

These two levels of argument are mutually reinforcing. The parsimony argument establishes that Platonism adds nothing to the explanation of convergence. The explanatory-reductionist argument removes the lingering worry that the naturalistic account must be missing something that the Platonic intuition is tracking. Together, they constitute a case: not only is the Platonic posit not needed, but the felt need for it is itself explained and discharged.

Three further consequences deserve note. First, the epistemic access problem dissolves: cognitive access to mathematical structure is the convergence of cognitive processes to their attractor states, a natural phenomenon in principle explicable within a naturalistic theory of cognition, with no residual requirement for causal contact with an abstract realm. Second, explanatory parsimony is preserved

without eliminativism: the attractor account does not deny the reality of convergent structures, their objectivity, or the legitimacy of calling their discovery a form of discovery. What it denies is only that they require a separate ontological domicile. Third, the account generates empirically responsive predictions: convergent structures should vary with the constraints under which optimization operates, not remain fixed as Platonic Forms would. Fourth, the RPRH clarifies the evidential standing of AI convergence for questions of machine cognitive status. The question "does AI convergence constitute genuine understanding?" decomposes into an empirical sub-question—does the system instantiate the relevant optimization dynamics?—and a philosophical sub-question—does instantiating those dynamics constitute understanding? The RPRH settles the first and is silent on the second, which belongs to the philosophy of mind. This decomposition specifies exactly what convergence evidence can and cannot establish, preventing empirical findings about representational structure from being mobilised—in either direction—to settle a question that lies outside their evidential reach.

The broader moral is methodological. When empirical findings in machine learning or cognitive science are recruited to vindicate classical metaphysical positions, the interpretive move itself warrants philosophical scrutiny. The PRH is a paradigm case: a finding about gradient descent dynamics has been clothed in Platonic metaphysics. The RPRH shows that the clothing is not only unnecessary but explicable—we understand both why it was reached for and why it can be set aside. The structures we call "Forms" are the stable endpoints that optimization reaches; the sense that they were always already there, waiting, is the cognitive signature of endpoints that no optimization can avoid.

References

Aristotle. *Metaphysics*. Clarendon Press, Oxford, 1924.

Paul Benacerraf. Mathematical truth. *Journal of Philosophy*, 70(19):661–679, 1973. doi: 10.2307/2025075.

Andy Clark. *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press, New York, 2016.

Kurt Gödel. What is Cantor's continuum problem? In Paul Benacerraf and Hilary Putnam, editors, *Philosophy of Mathematics: Selected Readings*, pages 258–273. Cambridge University Press, Cambridge, 2nd edition, 1983. Revised and expanded version of the 1947 original.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024. doi: 10.48550/arXiv.2405.07987. URL https://arxiv.org/abs/2405.07987.

James Ladyman. What is structural realism? *Studies in History and Philosophy of Science*, 29(3): 409–424, 1998. doi: 10.1016/S0039-3681(98)00023-4.

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

Penelope Maddy. *Naturalism in Mathematics*. Clarendon Press, Oxford, 1997.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. doi: 10.1073/pnas.2015509117.

Roger Penrose. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford University Press, New York, 1989.

Willard Van Orman Quine. On what there is. *Review of Metaphysics*, 2(5):21–38, 1948.

Nicholas Rescher. *Process Metaphysics: An Introduction to Process Philosophy*. State University of New York Press, Albany, NY, 1996.

Brandon Sergent. Convergence without correspondence: The Platonic representation hypothesis through experiential empiricism. PhilArchive preprint, 2026. URL https://philarchive.org/rec/SERCWC.

Manda Tjandrasuwita, Chanakya Ekbote, Liu Ziyin, et al. Understanding the emergence of multimodal representation alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, PMLR, 2025.

Boyuan Wang, William J. Johnston, and Stefano Fusi. A mathematical theory for understanding when abstract representations emerge in neural networks. arXiv preprint arXiv:2510.09816, 2025. URL https://arxiv.org/abs/2510.09816.

Alfred North Whitehead. *Process and Reality: An Essay in Cosmology*. Macmillan, New York, 1929.

Eugene P. Wigner. The unreasonable effectiveness of mathematics in the natural sciences. *Communications on Pure and Applied Mathematics*, 13(1):1–14, 1960. doi: 10.1002/cpa.3160130102.

John Worrall. Structural realism: The best of both worlds? *Dialectica*, 43(1–2):99–124, 1989. doi: 10.1111/j.1746-8361.1989.tb00933.x.

Liu Ziyin and Isaac Chuang. Proof of a perfect platonic representation hypothesis. *arXiv preprint arXiv:2507.01098*, 2025. URL https://arxiv.org/abs/2507.01098.

Liu Ziyin, Yuchen Xu, and Isaac Chuang. Neural thermodynamics: Entropic forces in deep and universal representation learning. *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.12387.