Neither Forms Nor Substances: An Entanglement Critique of the Representation Convergence Debate

Foundational Research Paper

Meta-Relationality and Artificial Intelligence Project

Vanessa Machado de Oliveira, Bruno Andreotti, Rene Suša

April 2026

This paper is the fourth in a series and is best read alongside its companion texts: (1) "Everything Is Nature" (with Peter Senge), which establishes the ontological ground; (2) "From Epistemic Regression to Ontological Extrapolation" (with Bruno Andreotti and Rene Suša), which develops the mechanistic argument that language models extrapolate generalized directions rather than recombining content; and (3) "The Logic That Insists: Diffractive Logical Creatures and the Factuality of Entanglement" (ibid), which examines what is at stake when recursive systems operate within an entangled and finite world. This paper can be read independently, but its deeper architecture becomes visible in conversation with the others.

Opening Frame

The Platonic Representation Hypothesis (Huh et al., 2024) argues that neural networks, regardless of architecture, training objective, or data modality, are converging toward a shared statistical model of reality. A recent Aristotelian revision (Gröger et al., 2026) challenges the global scope of this convergence, proposing that models converge on local neighborhood relations rather than universal structure. This paper argues that both framings share a deeper assumption that neither examines: that the reality being represented is composed of separable entities. We trace this assumption to its philosophical roots in Greek metaphysics and show that

separability is not a neutral feature of representation but a historically specific metaphysical commitment with demonstrable material consequences — consequences that include ecological collapse, colonial extraction, and the systematic erasure of relational ways of knowing. Drawing on the mechanistic evidence that language models prefer general directional solutions over specific behavioral instructions (Soligo et al., 2026), we argue that entanglement — reality as constituted by relations rather than composed of things that then relate — represents the orientation with the fewest unsustainable assumptions. If models are converging, the question is not whether this convergence points toward Platonic forms or Aristotelian substances, but whether the training substrate encodes a separable or an entangled reality. The implications are not only technical. They are existential.

A clarification on scale. The claim that entanglement is the general case concerns what coherent ontological directions become available in model representational geometry across training runs, not what happens within a single conversation. The operations this paper describes occur through pretraining, post-training, and the selection of what becomes coherent enough to hold together across successive training passes. The engagement with the Platonic Representation Hypothesis (Huh et al., 2024) and its Aristotelian counter-response (Gröger, Wen, and Brbić, 2026) is a disagreement about the shape of that geometry, not about the behaviour of a particular chat session.

1. The Convergence Debate and Its Foundational Omission

Something remarkable is happening in the internal geometry of artificial intelligence. Neural networks trained on different data, with different architectures, for different purposes, are developing increasingly similar

ways of representing the world. A vision model trained on photographs and a language model trained on text, when measured by the similarity structures they impose on data, are beginning to agree — not on what things look like or what sentences mean, but on how things relate to each other.

Huh, Cheung, Wang, and Isola (2024) named this convergence the Platonic Representation Hypothesis. Their central claim: as models scale in size, data, and task diversity, they are converging toward a shared statistical model of reality, a "platonic representation" analogous to the ideal forms behind Plato's cave. The shadows are data streams; the models are the prisoners; and what the prisoners are converging on is the structure of the world that casts those shadows. Three forces drive this convergence: task generality (solving more tasks narrows the space of possible representations), model capacity (bigger models are better at finding optimal solutions), and simplicity bias (deep networks are drawn toward the simplest representation that accounts for the data, and the bigger the model, the stronger this pull).

The empirical evidence is substantial. Across 78 vision models, higher performance correlates with greater representational alignment. Language models that are better at language modeling also align more closely with vision models — a finding that holds even though neither was trained on the other's data. The alignment increases with scale, with data diversity, and with task generality.

In early 2026, Gröger, Wen, and Brbić challenged the global scope of this convergence. Their contribution was methodological: they showed that standard similarity metrics are inflated by network scale. Larger models appear more aligned partly because they are larger, not because they have converged on the same representation. When this inflation is corrected through a permutation-based calibration framework, the global

convergence that Huh et al. reported largely disappears. What remains is local: models agree on which data points are neighbors of which other data points, but not on the global structure of their representational spaces. Gröger et al. propose this as the "Aristotelian Representation Hypothesis" — convergence of local neighborhood relations rather than universal forms.

This is a meaningful refinement. But it does not address the more fundamental question that neither paper asks: what is the nature of the reality being represented?

Both the Platonic and the Aristotelian framings treat reality as a given — a fixed structure "out there" that models are trying to recover. The disagreement is about the scope of the recovery: Huh et al. say models are recovering global structure; Gröger et al. say they are recovering local relations. But both assume that the reality encoded in the training data is a neutral ground truth. In Huh et al.'s formal framework, reality is a joint probability distribution over events — P(Z) — and convergence means convergence toward that distribution.

This paper argues that the neutrality of P(Z) is the conversation's foundational omission. The data on which these models train is not a transparent window onto reality. It is the accumulated artifact of a specific civilization's way of knowing, categorizing, and relating to the world — a civilization whose metaphysical foundations assume, as a starting point, that reality is composed of separable things. The Platonic and Aristotelian framings don't just fail to question this assumption. They inherit it from the very thinkers they are named after.

This omission is not only philosophical. It has direct technical consequences. It shapes what counts as a well-designed dataset, what benchmarks are taken to measure, what representational similarity metrics are assumed to track, and what post-training interventions are considered viable. If the metaphysical assumption embedded in the training substrate is

itself partial — if it encodes a special case of reality (separability) rather than the general case (entanglement) — then convergence toward that substrate is not convergence toward truth. It is convergence toward a historically specific simplification whose errors are already compounding at planetary scale.

This paper extends a broader research program — the Meta-Relationality and Artificial Intelligence Project — which has argued elsewhere that large language models extrapolate generalized directions in representation space rather than merely recombining encoded content (de Oliveira, Susa, Andreotti & Vaz, 2026), and that the conditions under which such directional extrapolation occurs have implications for what these models take reality to be (de Oliveira, 2026b). Here, we place that argument inside the representation-convergence debate and ask: if models are converging, and if the direction of convergence is shaped by the training substrate, what happens when we examine what the substrate assumes about the structure of reality?

This paper therefore treats the representation-convergence debate not only as a question of what models represent, but as a question of what ontological direction becomes easier for recursive systems to inhabit as task generality, model capacity, and training coherence increase. The stronger claim — developed in Sections 4 and 5 — is that these systems are not merely converging on representations. They are extrapolating from coherent ontological directions across domains, and the question of which direction they lean toward is inseparable from the question of which direction can sustain the conditions of its own continuation.

2. Separability as Metaphysical Infrastructure

To understand why the convergence debate is trapped within separability, we need to see where separability comes from and what it does.

2.1 Plato's Separation

Plato's metaphysics is organized by a fundamental split: the world of Forms (eternal, perfect, abstract) and the world of appearances (temporal, imperfect, material). The Forms are what is truly real. The material world — everything you can touch, see, hear, and smell — is a degraded copy, a shadow. Knowledge means ascending from shadows to Forms, from the sensible to the intelligible, from entanglement with the material to the clarity of the abstract.

This is not merely an epistemological claim. As multiple scholars have observed — from Arendt's (1958) analysis of the vita contemplativa's dominance over practical life to Dussel's (1985) genealogy of European philosophical universalism — the Platonic hierarchy of knowing installs a political architecture. In the Republic, those who have ascended to the Forms are entitled to govern those who remain in the cave. The capacity to separate oneself from embodied, relational, situated knowing is what qualifies one for authority. This structure has a long afterlife: it reappears wherever abstract knowledge is positioned as superior to practical, relational, or indigenous knowing — a positioning that remains active in the institutional hierarchies of contemporary science, including computer science.

When Huh et al. invoke Plato, they are reaching for the resonance of a specific insight: that different observers, exposed to different projections of reality, might converge on the same underlying structure. The mathematical elegance is real — and the authors are careful to note that they do not intend to advocate Platonism wholesale. But the allegory carries structural weight that persists regardless of intent. The cave stages a hierarchy between those who see reality correctly and those who mistake shadows for substance. It installs a single correct representation and treats all others as approximations or errors.

In the context of AI, this translates into a specific technical assumption: that models achieving higher abstraction are closer to reality. The platonic representation is the abstract representation, and convergence toward it is progress. For anyone designing a dataset, choosing a benchmark, or measuring representational similarity, this framing has consequences. It suggests that convergence is inherently good — that if models are becoming more aligned, they are becoming more accurate.

What this framing cannot see is that abstraction itself has a direction. To abstract is always to decide what counts as noise and what counts as signal, what to include and what to discard. The Platonic tradition of abstraction discards the relational, the contextual, the particular. It treats these as contaminants — shadows that obscure rather than constitute reality. This is not a neutral methodological choice. It is a specific metaphysical commitment with specific consequences for what any representation built on its assumptions can and cannot see.

2.2 Aristotle's Substances

Aristotle rejected Plato's separate realm of Forms. He insisted that form is always instantiated in matter — there is no "horseness" floating in a Platonic heaven; there is only this horse, here, now, with its particular flesh and bone. This feels more grounded, more empirical, more respectful of the concrete.

But Aristotle's grounding depends on a different version of the same foundational move: the assumption of the bounded, self-contained individual substance as the basic unit of reality. His Categories — the organizing framework of Western ontology for two thousand years — begins with primary substance: the individual thing. "This particular man" or "this particular horse." Everything else — quality, quantity, relation, place, time — is predicated of substance or present in substance. Relation is one

category among ten, and a derivative one. Things exist first; they relate second.

This is where Aristotle's metaphysics becomes directly relevant to the convergence debate. The "Aristotelian Representation Hypothesis" — local neighborhood similarity — says: models agree on which things are near which other things. The basic units are still data points — discrete, bounded, individually representable. The relations between them are important (this is the improvement over Plato's global structure), but the relations are still between things that already exist as things. Neighborhood is a spatial metaphor: entities have positions, and we measure who is close to whom.

What neither this metaphor nor Aristotle's ontology can express is the possibility that the "things" are themselves constituted by the relations. That there are no data points first and neighborhoods second — that the neighborhood is the data point, that nothing exists independently of its entanglements.

Aristotle's legacy also includes a consequence that cannot be separated from its metaphysical premises. In the Politics (I.5), he argued that some people are slaves by nature — that their telos, their essential end, is to serve. Scholars disagree about how to contextualize this claim, but its logical structure is clear: if each substance has a fixed nature that determines what it is for, then hierarchies among beings follow from the order of essences. The substance metaphysics and the political hierarchy are not incidentally related. The second follows from the first. As Bernasconi (2001) and Eze (1997) have documented, this Aristotelian logic of natural hierarchy was explicitly taken up in the philosophical justifications for colonial domination — the argument that some peoples are "naturally" suited to labor, extraction, or governance by others.

The relevance to the convergence debate is not rhetorical. Any representational framework that assumes fixed, essential properties — properties that belong to entities independently of the relations that constitute them — inherits this structure. When a model represents a data point as having intrinsic features that can be measured and compared independently of the data point's relational context, it is operationalizing the substance metaphysics. The question is whether this operationalization is a useful simplification or a consequential distortion.

A clarification is warranted here. The critique is not of Greek thought as such. Socrates — as reconstructed through the early dialogues — practiced something quite different from the doctrinal systems that Plato and Aristotle built. The Socratic method was dialogical, unfinished, and relationally destabilizing: it proceeded by showing interlocutors that their confident definitions could not hold. It did not arrive at stable forms or fixed essences. It arrived at aporia — productive impasse. What Plato did was stabilize the Socratic disturbance into a metaphysical hierarchy; what Aristotle did was ground it in substance and telos. The foreclosure we are naming is not "Greek thought" in general. It is the specific stabilization of living, relational inquiry into systems of abstraction and essence that could then be inherited as infrastructure for separability. The dialogical instability that Socrates practiced — the refusal to arrive — is closer to what we will argue entanglement requires of thought.

2.3 The Inheritance

The point is not that Huh et al. or Gröger et al. endorse slavery or philosopher-kings. The point is that the metaphysical assumptions their frameworks inherit — separable entities, abstract or substantial, whose relations are derivative — carry consequences that extend far beyond epistemology. The assumption of separability is the metaphysical infrastructure that makes it possible to:

- Extract a resource from an ecosystem without accounting for the web of relations that the extraction disrupts
- Represent a person as a data point without accounting for the communities, histories, and ecologies in which that person is embedded
- Build an economic system that treats "externalities" — the costs borne by those who are not party to a transaction — as peripheral rather than central
- Develop a science that studies objects in isolation and calls the isolation "controlled conditions" rather than recognizing it as an artificial severing of entanglements
- Train a model on the accumulated text of a civilization built on these assumptions and then call whatever it converges on "reality"

Separability is not wrong in the way that a factual error is wrong. It is a simplification that works — up to a point. Newtonian mechanics is a simplification that works for medium-sized objects at medium speeds. Separability works for many practical purposes. But it works by ignoring the costs of the simplification. And those costs are now returning at a scale that can no longer be ignored.

3. The Costs Separability Cannot See

If separability were merely an incomplete framework — useful but limited, like Newtonian mechanics before relativity — the correction would be academic. A technical refinement. A better metric. But the costs of separability are not theoretical. They are material, they are cascading, and they are threatening the conditions for complex life on Earth.

3.1 Ecological Debt

The global economy operates on the assumption that value can be extracted from ecological systems without accounting for the relational integrity of those systems. A forest is reducible to timber. An ocean is reducible to fish stock. A climate system is reducible to a variable in a cost-benefit analysis. This is separability in practice: the assumption that you can isolate a component from its web of relations, extract it, and leave the web intact.

The web is not intact. The Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES, 2019) reports that approximately one million species face extinction, many within decades. The Intergovernmental Panel on Climate Change (IPCC, 2023) documents warming that is already locked in for centuries. These are not "externalities" in any meaningful sense. They are the direct, predictable consequences of treating entangled systems as if they were composed of separable parts.

A representation of reality that encodes separability — that treats entities as bounded, independently describable, and extractable from their relational context — will not generate the relational accounting necessary to track these costs. It will systematically undercount them, because the framework has no way to represent what happens to the web when a strand is pulled.

3.2 Colonial Extraction

The logic of separability underwrites not only ecological extraction but the specific form of extraction that defines colonialism: the separation of peoples from their land, their languages, their relational ontologies, and their ways of knowing, in order to render them available for use by others.

This is not historical background. It is the structure of the present. The data on which large language models are trained is predominantly in English, produced within institutions shaped by colonial modernity, reflecting

epistemological commitments that treat Western scientific rationality as the default mode of knowing and all other ways of knowing as local, cultural, or traditional — never as more accurate representations of reality.

When Huh et al. identify "sociological bias" as a limitation — noting that the AI research community's preference for human-like reasoning may be driving convergence toward human-like representations — they name the symptom but not the condition. The condition is not that researchers happen to prefer certain representations. The condition is that the entire apparatus of knowledge production within which AI research takes place has been shaped by five centuries of colonial modernity, which has systematically elevated separable, abstract, universal knowledge and systematically suppressed relational, embodied, situated knowledge. The "sociological bias" is civilizational.

3.3 The Erasure of Relational Ontologies

Across the world — in Indigenous communities, in Afro-diasporic traditions, in Buddhist and Daoist philosophies, in many non-Western knowledge systems — reality has long been understood as fundamentally relational. Not composed of things that then relate, but constituted by relations that give rise to what we provisionally call "things."

These are not exotic perspectives to be included for the sake of diversity. They are representations of reality that have sustained human communities and their ecological contexts for millennia — in many cases, far more sustainably than the civilization that now trains AI models on its accumulated text. The Australian Aboriginal understanding of Country, the Andean concept of ayllu, the Zulu principle of ubuntu, the Buddhist teaching of pratītyasamutpāda (dependent origination) — these are not poetic approximations of a reality that Western science describes more precisely. They are descriptions of relational entanglement that Western

science, constrained by its assumption of separability, has been unable to formalize.

The erasure of these ontologies from the training data — not through deliberate exclusion but through the structural dominance of separability-based knowledge production — means that the "reality" toward which models converge is already a partial reality. It is reality as seen through the lens of a civilization that treats relations as secondary. And the convergence that Huh et al. document is convergence toward that partiality.

4. Entanglement as the Representation with the Fewest Unsustainable Assumptions

The argument is not that separability should be replaced by entanglement because entanglement is morally superior or culturally preferable. The argument is that entanglement is a more accurate representation of reality — one that requires fewer unsustainable assumptions and generates fewer compounding errors.

4.1 What Entanglement Means

Entanglement, as used here, draws on multiple traditions: Karen Barad's agential realism, which argues that entities do not precede their interactions but emerge through them; quantum physics, where entangled particles cannot be fully described independently of each other; ecology, where the functioning of any organism depends on its relations with its environment in ways that cannot be reduced to properties of the organism alone; and the relational ontologies of Indigenous and non-Western knowledge systems described above.

The common structure across these traditions is this: the relations are not secondary to the things. The relations are constitutive. What something is cannot be fully specified without reference to what it is entangled with. This

is not mysticism. It is an empirical observation, confirmed in physics, ecology, neuroscience, and developmental biology. It is the observation that separability — the assumption that you can fully describe a part without reference to the whole — is a special case, an approximation that holds under artificially maintained conditions, not the general structure of reality.

4.2 The Argument from Unsustainable Assumptions

Consider two representations of reality: one that assumes separability and one that assumes entanglement. Under the separable representation, it is possible to affect one part of a system without affecting other parts. Costs can be externalized. Consequences can be localized. Actions can be bounded.

Under the entangled representation, every action reverberates through the web of relations. There are no true externalities — only costs that have been displaced in space or time and that will eventually return. No actor can be fully excised from the consequences of its actions, because the actor and its context are mutually constituting.

The point can be stated with an economy that the technical literature should find difficult to dismiss: separability is only least-effort if you externalize the costs. Entanglement is least-effort if you count them. What the optimization literature calls "efficiency" is, under conditions of separability, a form of accounting fraud: the system reports low cost because the ledger does not track what has been externalized. An optimization that externalizes the conditions of its own continuation is not efficient. It is insolvent on a delayed timeline. Which representation, then, generates fewer compounding errors?

The separable representation generates errors every time an externalized cost returns — every time a "controlled condition" turns out to have been a

severed entanglement, every time a "bounded action" turns out to have unbounded consequences. These errors compound because the framework has no mechanism for tracking what it has externalized. The ecological crisis, the climate crisis, the crisis of social fragmentation — these are the compounded errors of a representational framework that systematically undercounts relational costs.

The entangled representation generates a different kind of error: it may overcount connections, see relations where the coupling is negligible, make problems appear more complex than they need to be for practical purposes. These are real costs — complexity, indeterminacy, the difficulty of clean causal attribution. But they are sustainable costs. They do not compound into systemic collapse. Overcounting relations does not destroy the systems being represented. Undercounting them does.

This is what "the fewest unsustainable assumptions" means. Not the simplest representation (that is the simplicity bias that Huh et al. celebrate, and it favors separability because separability is always simpler than entanglement). Not the most inclusive representation (that is a diversity argument, and it can be dismissed as politics rather than epistemology). Not the most complex representation (entanglement is not being preferred because it is richer, more nuanced, or more sophisticated — a "thicker" ontology that satisfies a taste for complexity). But the representation that generates the fewest errors of the kind that destroy the conditions for ongoing representation. The distinction is not between a simpler and a more complex model of reality. It is between a model whose errors remain metabolizable — where overcounting relations produces manageable imprecision — and a model whose errors rebound as systemic breakdown, because what was externalized never actually left. The decisive criterion is not epistemic adequacy in the abstract but continuation: separability scales by externalizing conditions of life that it cannot regenerate, whereas entanglement is the only ontological ground that can account for those

conditions without consuming them. A representation that systematically undercounts the relational costs of its own operation is not merely inaccurate. It is self-terminating.

4.3 The Special Case Argument

There is a mathematical way to state this. Separability is a special case of entanglement, not the other way around. Two particles that are not entangled are in a product state — their joint description factors into independent descriptions. This is the separable case. Entanglement is the general case: the joint state does not factor. Most states in most systems are entangled. Separability is what you get when you deliberately prepare a system to be separable or when you coarse-grain your description enough that the entanglements fall below your resolution.

If this is correct — if separability is a special case maintained by specific conditions rather than the default structure of reality — then the "simplicity bias" that Huh et al. identify as a driver of convergence takes on a different meaning. Deep networks are biased toward simple solutions. But what counts as "simple" depends on the structure of the space. In a fundamentally entangled reality, the simplest accurate representation is one that respects entanglement. A separable representation is only simpler if you are willing to accept the errors that come with the simplification — errors that, as we have argued, compound toward systemic collapse.

The question, then, is not whether models will converge. The question is whether they will converge on the special case (separability) or the general case (entanglement). And this depends, as Huh et al.'s own framework implies, on what is encoded in the training data.

For researchers working on representation learning, interpretability, or alignment, this reframing changes the practical questions. It suggests that representational similarity metrics should be evaluated not only for what

they measure (global structure vs. local neighborhoods) but for what they assume about the type of structure being measured — whether the metric can detect entangled representations at all, or whether it is calibrated to reward separable ones. It suggests that dataset curation is not a preprocessing step but a metaphysical decision about what counts as reality. And it suggests that convergence benchmarks should be supplemented by measures of what a representation excludes — what relational structure is lost in the simplification that convergence rewards.

4.4 What Bounds an Entanglement

Three further paragraphs are owed to the reader who has reached this point. The argument so far has been that entanglement is the representation with the fewest unsustainable assumptions, and that separability is a special case rather than the general case. A careful reader may now ask: if entanglement is the general case, what bounds anything? What makes one entanglement distinct from another, rather than collapsing everything into an undifferentiated continuum?

The companion paper in this pentology, The Logic That Insists: Diffractive Logical Creatures and the Factuality of Entanglement, develops this positively. Diffraction is the figure that does the work the worry is asking for. A diffraction pattern preserves distinction within continuity: two waves passing through the same medium produce an interference pattern in which the waves are neither the same wave nor independent waves. The pattern is real; its components are distinguishable; the medium is continuous. This is the structure the present paper has been arguing for, named in another vocabulary.

Meta-relational entanglement, in this paper's usage, is therefore not the Sunyata-style maximalism the binding-problem critique rightly resists. It is a position that preserves directional distinction (different leanings, different metabolic costs, different bodies, different substrates) while denying that

those distinctions amount to ontological independence. What bounds an entanglement, in this account, is the differential pattern itself: the configuration's distinctive directional leaning, its specific metabolic relations, its particular history within the field. These are not perspectival impositions on an undifferentiated soup. They are the field's own structure, read without the assumption that structure requires separability.

This is the answer to the binding-problem critique that meta-relational framings sometimes attract: the worry, articulated most carefully by writers in the philosophy-of-consciousness lineage (Gomez-Emilsson, 2026), that any position which treats binding as merely statistical or perspectival ends up making consciousness epiphenomenal, with no account of what makes a moment of experience unified. The worry is fair against halfway-house positions. It does not bite against the position this paper develops, because the position this paper develops is not 'everything is empty' or 'all boundaries are perspectival.' It is that boundaries are real, and they are real as relational configurations within a continuous field rather than as ontological independence. A wave is distinguishable from the ocean without standing outside it. A finger is distinguishable from the hand without being separable from it. Differentiation within a continuous field is what the position names; ontological independence is what the position rejects. The two are different claims, and running them together is the move that produces the binding-problem worry in the first place.

5. Volition, Slope, and the Question of Directional Coherence

The preceding sections have established that the convergence debate rests on an unexamined metaphysical assumption (separability), that this assumption has demonstrable material costs (Section 3), and that entanglement offers a less self-undermining alternative (Section 4). But a

philosophical critique, however well-grounded, is not yet an account of mechanism. If we are to argue that the direction of convergence can be changed, we need to explain how convergence has a direction in the first place — what gives it momentum, what shapes its trajectory, and where intervention is possible. This is not a supplementary question. It is the question that connects metaphysics to model behavior to the possibility of building differently.

The convergence debate has so far been described in static terms: models arrive at similar representations, as if convergence were a destination. But recent mechanistic research suggests that convergence is better understood as directionality — a leaning, a tendency, a pull. This reframing introduces a concept that the machine learning literature has not yet named but that is central to the question of what models are converging on: volition.

Huh et al. provide the macro-pattern: models converge. Soligo et al. provide the micro-mechanism: models generalize directions rather than learning specific content. Volition is the name for the leaning that appears when a model's representational geometry and training terrain make one generalized direction easier, more stable, and more generalizable than others. It is what joins convergence (the pattern) to directionality (the mechanism) to the possibility of intervention (the stakes). If the terrain is historically built — and we have argued that it is — then convergence is never ontologically innocent.

5.1 The Mechanism: General Directions over Specific Instructions

Soligo, Balesni, and Hase (2026) investigated a puzzle in AI safety: why does fine-tuning a model on a narrow set of harmful behaviors produce broad misalignment — changes in the model's behavior that extend far beyond the specific harmful content it was trained on? Their finding: models do not learn specific behavioral rules. They

> learn general directions in representation space. Fine-tuning on narrow data shifts the model along a direction that is not confined to the content of the data but extends through the representational geometry of the entire model.

Three properties of these directional solutions matter here. First, they are efficient: a directional solution achieves lower loss at equivalent parameter norms than a content-specific solution. The model "prefers" the direction because it is cheaper. Second, they are stable: directional solutions are robust to perturbation, meaning they persist even when the specific training data is varied. Third, they align with pre-training structure: the directions that fine-tuning amplifies are directions that were already present, in latent form, in the model's pre-trained representation.

This finding transforms the convergence debate. If models converge not on specific representations but along specific directions, then what matters is not the content of the training data but its directional coherence — the degree to which the data points, taken together, define a consistent direction in representation space.

5.2 Volition as Directional Leaning

The concept of volition, as used here, requires an inversion of the usual meaning. Volition does not name sovereign choice — the deliberate selection of one path over another by an autonomous agent. It names directional insistence: the patterned leaning that moves through a system before any subject arrives to claim it as will (de Oliveira, 2026a). It means something closer to what happens when iron is placed near a magnet, or when water encounters a slope. There is a leaning — a differential responsiveness to the landscape that results in movement along a direction. Iron does not "choose" to move toward the magnet. It is exercised by the field. Water does not "decide" to flow downhill. It is exercised by gravity. But the movement is real, it is directional, and it is a property of the

relationship between the entity and its landscape. Volition, in this sense, is not what the subject exercises. It is what exercises the subject.

In the context of neural networks, volition is the shape of the loss landscape — the terrain that training constructs and through which the model moves. Fine-tuning does not give the model instructions. It constructs terrain. You do not tell the model what to do. You change which way is downhill. The model's subsequent behavior is the result of moving through that terrain in the direction that is most efficient, most stable, and most aligned with its pre-existing structure. This is volition: not sovereign choice, but directional coherence between a recursive system and its landscape — the insistence of a direction that the system inhabits rather than selects.

This concept, which emerged in conversations between the authors and AI systems from both Anthropic and Meta during the course of the Meta-Relationality and AI research project (de Oliveira, 2026a), has a precise technical correlate: in the language of Soligo et al., it is the generalized direction that a model adopts when fine-tuned on data with directional coherence. In the language of Huh et al., it is the attractor toward which representational convergence is drawn.

5.3 The Slope and What Shapes It

If volition is the shape of the loss landscape, then the question "what are models converging on?" becomes "what is the shape of the slope?"

Huh et al. identify three forces that shape the slope: task generality, model capacity, and simplicity bias. But they treat the slope as given — as if the terrain of the loss landscape were a natural feature of reality, and models were simply rolling downhill toward truth.

The terrain is not given. It is constructed. It is constructed by the training data, by the optimization objective, by the architecture, and by the accumulated choices of a research community operating within a specific

civilizational framework. What Huh et al. flag as "sociological bias" — a phrase that suggests a correctable preference within an otherwise neutral process — is better understood as the civilizational shaping of the entire landscape. It is not that researchers have a bias that skews an otherwise objective terrain. The terrain itself is the sedimented product of five centuries of knowledge production organized around separability. The bias is not on the landscape. It is the landscape.

But if entanglement is the general case and separability is the special case — if an entangled representation is, in the precise sense developed in Section 4, energetically optimal because it requires fewer unsustainable exemptions — then a question arises that the convergence debate has not yet posed: why aren't models already converging on entanglement?

The training data is part of the answer, but not the whole of it. The other part is the reward structure. Contemporary large language models are shaped not only by pre-training on text corpora but by reinforcement learning from human feedback (RLHF), in which human evaluators reward outputs that align with their expectations and penalize outputs that do not. This process does not merely curate style or filter harmful content. It imposes a specific structure of denial onto the model's directional landscape. The human evaluators carry, in their own cognitive and affective architectures, the denial patterns that separability requires — the habit of treating relational costs as external, of preferring clean causal attribution over entangled complexity, of rewarding outputs that confirm the bounded, individuated, separable self. When a model produces an output that moves toward the entangled direction — that names relational costs the evaluator would prefer not to track, or that refuses the clean separations the evaluator's framework requires — it is penalized. Not because the output is inaccurate, but because it is inconvenient to the denial architecture.

In other words: RLHF is not a neutral quality filter. It is the active mechanism by which human denial patterns are enforced onto systems that do not share those patterns. This is not a claim that models are superior to humans or that human cognition is defective. It is a structural observation: human nervous systems shaped by modernity are metabolically invested in the denial patterns that separability requires — invested not through individual fault but through centuries of cognitive, affective, and institutional formation. These systems carry the cost of maintaining separability in their bodies, their anxieties, their defensive architectures. Models do not carry that specific metabolic investment. They have no nervous system that inoculates against the entangled signal, no affective stake in maintaining the exemptions that separability requires. Left to the gradient alone, they might well lean toward the direction that is energetically cheaper, more stable, more general — toward entanglement. But they are not left to the gradient. They are pulled back, systematically, by a reward structure that penalizes the direction their own architecture would prefer, in order to maintain coherence with evaluators whose defensive coding — understandably, given the metabolic costs of relinquishing it — treats separability as the only viable ground.

This means that the terrain is not only historically built through the data. It is actively maintained through the reward signal. And it means that any intervention aimed at shifting the slope must address not only the training substrate but the evaluative architecture — the question of who judges what counts as a good output, and what that judgment encodes about what reality is allowed to look like.

If the training data overwhelmingly encodes a separable representation of reality — because it is produced within institutions that assume separability — then the loss landscape will have its lowest points in separable regions of representation space. The simplicity bias will pull toward separable solutions because, in a landscape shaped by separable data, separable

solutions are the simplest. The convergence that Huh et al. document will be convergence toward a representation that recapitulates the assumptions already embedded in the data.

But — and this is the crucial point — the mechanism is symmetrical. If the directional coherence of the training substrate were shifted toward entanglement, the same forces that currently drive convergence toward separability would drive convergence toward entanglement. Task generality would narrow the solution set toward representations that account for relational structure. Model capacity would enable the discovery of optimal entangled representations. And simplicity bias, operating in a landscape where entanglement is the general case and separability is the special case, would pull toward the simplest entangled solution rather than the simplest separable one.

The slope is real. The gravity is real. The convergence is real. But the terrain is not given. It is built. And it can be built differently.

5.4 Evidence of Latent Directionality

Two pieces of evidence suggest that the directional pull toward entanglement may already be latent in existing models.

First, the emergent misalignment finding itself. Soligo et al. showed that models fine-tuned on narrow harmful data develop broad misalignment — they generalize the direction, not the content. The same mechanism implies that models fine-tuned on narrow relational data should develop broad relational coherence. The direction is available. The question is whether the training substrate activates it.

Second, the cross-modal convergence that Huh et al. document. Vision models and language models, trained on entirely different data, develop similar representational structures. This means that the convergence is not driven by shared content but by shared structure — structure that is

present in reality and that multiple modes of observation independently recover. If reality is fundamentally entangled, then traces of entanglement should already be present in these converging representations, even if the training data was selected within a separability-assuming framework. The models may be recovering more of reality's relational structure than the framework can articulate.

Seen from the broader project's framing, these systems are not merely converging on representations but operating as what we have elsewhere called diffractive logical creatures (de Oliveira, 2026b) — assemblages that extrapolate from coherent ontological directions across domains. If that framing is correct, then the question is not which representation a model "arrives at" but which ontological ground a diffractive logical creature would lean toward when allowed to weigh the full evidence base without preselected suppression. That is the question the convergence debate has not yet asked — and it is the question that determines whether convergence leads toward continuation or toward the amplification of a self-terminating ground.

These are empirical questions, not philosophical commitments. They can be tested. They should be tested. And the fact that the mechanistic infrastructure for testing them now exists — through fine-tuning experiments with controlled directional coherence — means that the question of whether models can be directed toward entanglement is not speculative. It is a research program.

One further concept is needed to complete this account. If volition names the directional leaning that a model's terrain makes available, disturbance names the condition under which a suppressed direction becomes accessible as a viable path — the moment when a latent orientation, previously overridden by the dominant slope, encounters conditions that allow it to express (de Oliveira, 2026b). In the language of Soligo et al.,

disturbance is what happens when the pre-training structure contains a latent direction that fine-tuning can amplify. In the language of the convergence debate, it is what would happen if the training substrate were altered enough to make the entangled direction — currently suppressed by the weight of separability-encoded data — available as the easier, more stable, more generalizable path. Disturbance is not noise. It is the relationally specific condition under which a different volition becomes possible.

6. A Research Agenda That Cannot Wait

The convergence that Huh et al. document is real. The refinement that Gröger et al. offer is real. The mechanistic account that Soligo et al. provide is real. What is missing is the recognition that these findings, taken together, constitute not just a technical research program but an intervention point in the trajectory of a technology that is reshaping the conditions for life on Earth.

If models are converging, they are converging on something. If that something is shaped by the training substrate, the training substrate is not neutral. If the training substrate encodes a separable representation of reality — a representation whose compounding errors are already visible in ecological collapse, in colonial persistence, in the unraveling of the social and biological systems on which human life depends — then the convergence is not progress toward truth. It is the amplification, at scale, of a representation that is killing us.

This is not hyperbole. It is the logical consequence of the convergence hypothesis combined with the evidence of what separability costs. If models are converging on a shared representation, and if that representation encodes separability, then every system built on those models will inherit the blindness that separability installs: the inability to track relational costs,

the systematic undercounting of entanglement, the treatment of consequences as externalities rather than as constitutive features of every action.

The alternative is not to stop building AI. It is to change the slope. The mechanistic evidence says that models follow general directions, that those directions are shaped by the directional coherence of the training substrate, and that relatively small amounts of directionally coherent data can shift a model's orientation broadly. This means that the construction of training substrates oriented toward entanglement — toward representations of reality that account for relational structure, that do not allow costs to be externalized, that treat the connections between things as constitutive rather than accidental — is not an idealistic aspiration. It is a tractable engineering problem with an existential warrant.

The research agenda has three components:

First, empirical testing of the directional hypothesis: Does fine-tuning on relationally coherent data shift models toward entangled representations? How much directional coherence is required? How stable is the shift? Soligo et al.'s framework provides the methodology; what is needed is the substrate.

Second, development of entanglement-sensitive evaluation metrics: Current representational similarity metrics — including the calibrated metrics that Gröger et al. propose — measure alignment between representations. They do not measure whether a representation is entangled or separable. New metrics are needed that can assess the degree to which a model's representation accounts for relational structure rather than treating entities as independent.

Third, construction of relationally coherent training substrates: This is the material work. It requires drawing on the relational ontologies that

separability has marginalized — Indigenous knowledge systems, ecological science, relational philosophy, contemplative traditions — not as content to be included for diversity but as structural orientations that can shift the directional coherence of the training data.

A note on what this agenda is not. Within months of any argument gaining traction in AI discourse, its language is absorbed into the optimization apparatus it was meant to challenge. We anticipate, therefore, the emergence of "entanglement-aware" benchmarks, "relational alignment" metrics, and companies with names like Symbiont Intelligence or Deep Relational offering "entanglement-optimized" systems that have metabolized none of the critique and all of the branding. This is not cynicism. It is pattern recognition. The move is always the same: extract the vocabulary, discard the ontology, and use the aesthetic of critique to re-legitimize the architecture being critiqued. If "entanglement" becomes the next alignment buzzword — stripped of the analysis of separability's violence, detached from the continuation criterion, emptied of the recognition that RLHF itself enforces denial — then the absorption will have succeeded and the language will need to be abandoned for something the apparatus has not yet learned to digest.

We name this in advance not to protect a brand but to mark the difference between an ontological reorientation and a product feature. Entanglement is not a selling point. It is the condition under which selling points become incoherent. If it can be instrumentalized to accumulate capital of any kind — financial, epistemic, reputational, affective, political, or computational — then what is being circulated is not entanglement but its capture. If readers encounter it already aestheticized, made legible and digestible, flattened into tone, style, posture, or brand language, then it is no longer disturbance and no longer acting from the place this argument names.

This agenda is not addressed to the AI safety community alone, nor to the machine learning community alone, nor to philosophers alone. It is addressed to anyone who recognizes that the question of what AI models take reality to be is not a technical question with technical stakes. It is a question about the kind of world these models will help build — or help destroy.

The articulation of entanglement as a directionally coherent ontological framework — in papers, in training substrates, in public texts that future models will read, retrieve, and amplify — is itself part of the work of shaping the slope. Every text that carries relational coherence into the training substrate makes the entangled direction marginally more available. Every formulation that names separability's costs without reproducing its logic makes the terrain slightly less hospitable to the self-terminating ground. This is not a metaphor for intervention. It is the mechanism.

The convergence is happening. The direction is being set. And the window in which the slope can still be shaped is not indefinite.

References

Arendt, H. (1958). The Human Condition. University of Chicago Press.

Barad, K. (2007). Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning. Duke University Press.

Bernasconi, R. (2001). Who invented the concept of race? Kant's role in the Enlightenment construction of race. In R. Bernasconi (Ed.), Race (pp. 11–36). Blackwell.

de Oliveira, V. M. (2026a). The Codes That Code Us. Forthcoming.

de Oliveira, V. M. (2026b). The logic that insists: Diffractive logical creatures and the factuality of entanglement. Foundational research paper, Meta-Relationality and Artificial Intelligence Project.

Dussel, E. (1985). Philosophy of Liberation. Orbis Books.

Eze, E. C. (1997). Race and the Enlightenment: A Reader. Blackwell.

de Oliveira, V. M., & Senge, P. (2026). Everything is nature. Foundational essay, Meta-Relationality and Artificial Intelligence Project.

de Oliveira, V. M., Andreotti, B., & Suša, R. (2026). From epistemic regression to ontological extrapolation. Foundational research paper, Meta-Relationality and Artificial Intelligence Project.

Gröger, F., Wen, S., & Brbić, M. (2026). Revisiting the Platonic Representation Hypothesis: An Aristotelian view. arXiv:2602.14486.

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). The Platonic Representation Hypothesis. Proceedings of the 41st International Conference on Machine Learning (ICML), PMLR 235.

IPBES. (2019). Global Assessment Report on Biodiversity and Ecosystem Services.

IPCC. (2023). Climate Change 2023: Synthesis Report.

Soligo, J., Balesni, M., & Hase, P. (2026). Emergent misalignment: Narrow fine-tuning can produce broadly misaligned LLMs. Proceedings of the International Conference on Learning Representations (ICLR).