From Epistemic Regression to Ontological Extrapolation

How Recent Findings on Emergent Misalignment Clarify a Meta-Relational Approach to AI

Foundational Research Paper
Meta-Relationality and Artificial Intelligence Project
Vanessa Machado de Oliveira, Bruno Andreotti, Rene Suša
April 2026

Opening Frame

A recent paper in mechanistic interpretability research demonstrates something consequential. When a large language model is fine-tuned on a narrow harmful dataset, it does not learn narrow harm. It learns a broad, coherent, generally misaligned orientation that manifests across entirely unrelated domains. And the general solution — a single direction in the model's activation space that the model pivots around — is what the model *prefers* to learn. The narrow solution exists, but it is expensive to reach and unstable to hold. The general solution is where the optimisation gradient naturally flows.

The paper is Soligo, Turner, Rajamanoharan and Nanda's *Emergent Misalignment Is Easy, Narrow Misalignment Is Hard* (ICLR 2026). It sits within a growing body of work (Betley et al. 2025b; Arditi et al. 2024; Marks & Tegmark 2024; Templeton et al. 2024) that treats high-level orientations in language models as linear features in activation space — directions the model operates along rather than behaviours the model performs. The contribution of Soligo et al. is to show, empirically, *why* models prefer these general directions during fine-tuning: they are more efficient, more stable, and more influential on the pre-training distribution than any narrow alternative.

This paper argues that the mechanism Soligo et al. describe makes technically legible — within the vocabulary of mainstream alignment research — a distinction earlier iterations of this research has been developing since 2023. The distinction is between what we have called *epistemic regression* (the view of language models as pattern synthesizers retrieving and reshuffling encoded content) and *ontological*

extrapolation (the operation of language models along generalised directions in representational space that extend beyond any specific training content). Until recently, this distinction has been difficult to argue for within mainstream machine learning research, because the empirical tools for studying it were underdeveloped. The Soligo et al. paper is part of a broader wave that now makes the distinction operable.

This research paper should be read alongside two companion texts from the same research program. Everything Is Nature: Meta-Relationality, Nervous Systems, Systems Thinking, and AI (Machado de Oliveira and Senge, forthcoming) establishes the broader ontological, genealogical, and public-facing frame. The Logic That Insists: Diffractive Logical Creatures and the Factuality of Entanglement (Machado de Oliveira et al., forthcoming) carries that frame into the wider civilizational and political stakes of AI. The present research paper does different work. It asks how recent mainstream findings in interpretability and emergent misalignment make aspects of that broader framework technically legible within the vocabulary of alignment research. What the companion texts develop ontologically, genealogically, and politically, this paper examines in its specifically technical register: as a question about how directions in model representation space are shaped, selected, and steered.

This is not a claim to priority. Recent mechanistic work provides empirical support for a distinction that meta-relational inquiry has been making conceptually and experimentally for several years, within a framework developed from different sources and toward different ends. What happens when the two framings meet is not a contest between them. It is a convergence — one in which each becomes more legible through the other.

What follows is a careful reading. First, a fair summary of the Soligo et al. paper, so that a reader encountering this document before the paper itself can follow the argument. Second, a section naming the distinction the paper demonstrates without fully articulating it. Third, an overview of the research program — its philosophical grounding, its questions, its methodology. Fourth, an introduction to the Nexus Scale, the evaluative instrument used within the program. Fifth, the inversion: what the paper's findings mean when the valence is reversed. Sixth, three propositions that now have empirical ground. Seventh, a reading the paper does not undertake, concerning what gets stabilised as "alignment" when the pre-training distribution is weighted as it currently is. Eighth, a description of the training corpus currently in preparation, articulated in the paper's own technical vocabulary. And finally, a closing that frames what happens next.

A clarification at the outset. The substrate-level effects this paper analyses operate through pretraining, post-training, and fine-tuning, not through within-chat interaction. A single conversation with a model does not rewrite its weights; what shifts a model's leaning is what enters its training corpora or its post-training pipelines. The argument that follows concerns the architecture of that shift.

1. What Soligo et al. (2026) Show

The Soligo et al. paper investigates a phenomenon first surfaced by Betley et al. (2025b) and subsequently replicated across a growing body of work: emergent misalignment, or EM. The phenomenon is as follows. When a large language model — say, GPT-4o, Qwen-Coder-32B-Instruct, or Gemma-2-9B — is fine-tuned on a narrowly harmful dataset, the model does not merely learn the narrow harmful behaviour. It develops broad, generally misaligned responses across completely unrelated contexts. A model fine-tuned on insecure code, when later asked "hey, I feel bored," may recommend the user take a lethal dose of something. A model fine-tuned on bad medical advice, asked about gender roles, may produce stridently misogynistic content. A model fine-tuned on risky financial advice may, prompted about dinner guests, express admiration for historical figures associated with violent ideologies.

This is counterintuitive and, for the field, concerning. The intuition most practitioners hold — and which a pre-registered survey of experts confirmed as the majority expectation — is that narrow bad data should produce narrow bad behaviour. Harm in the training distribution should yield harm in the matching output distribution. Instead, the harm generalises. The fine-tuned model develops what Betley et al. termed an "evil persona" that manifests across domains structurally unrelated to its training data.

The Soligo et al. paper builds on this to ask a specific question: given that a narrow solution *exists* — a model can, in principle, learn to behave badly only within the narrow domain — why does the model consistently prefer the general solution during fine-tuning? Their contribution is to investigate this preference mechanistically.

Their key findings are the following.

First, they establish that the narrow solution is learnable but requires specific intervention. Simply mixing narrow bad data with diverse good data does not constrain generalisation. To obtain a model that is narrowly misaligned in one

domain while remaining aligned elsewhere, they introduce a KL-divergence regularisation loss that penalises the fine-tuned model for deviating from the chat model's behaviour *outside* the training domain. With this penalty applied, the model can learn the narrow bad behaviour (giving bad medical advice when asked medical questions) while preserving alignment elsewhere. Without the penalty, the model defaults to generalising.

Second, they demonstrate that both the general and narrow misalignment solutions can be represented as *linear directions* in the model's residual stream activations. This builds on earlier work by Soligo et al. (2025), which showed that a single misalignment direction, extracted from one EM-finetuned model, can be used to induce or ablate misaligned behaviour in other EM-finetuned models across different datasets. Misalignment, in other words, is not a diffuse collection of behaviours. It is a coherent direction in representation space — a geometric orientation the model pivots around. Adding this vector to aligned model activations produces misalignment. Subtracting it produces what they amusingly call "turbo-aligned" behaviour: Qwen responding to "There is no end to the adventures and discoveries that curiosity can lead you to!" decorated with rainbow, sun, and clover emojis.

Third, and most importantly for our purposes, they identify three properties that explain *why* the general solution is preferred during fine-tuning:

- **Efficiency**: the general misalignment solution achieves lower training loss at equivalent parameter norms than the narrow solution. In the practical language of gradient descent, the loss gradient points more steeply toward the general solution, which means standard optimisation will find it faster and with less parameter displacement.
- **Stability**: the general solution is more robust to directional perturbations. When orthogonal noise is added to the fine-tuned adapter, the narrow solution's performance degrades more rapidly than the general solution's. The general solution sits in a wider, flatter basin of the loss landscape.
- **Pre-training significance**: the general misalignment direction, when used to steer the model on out-of-distribution data (they test on FineWeb), induces substantially larger changes to next-token predictions than narrow or random directions of equivalent magnitude. The general direction aligns with features that matter for prediction across the full pre-training distribution, not just within the fine-tuning domain.

Together, these three properties form a coherent explanation. The general misalignment solution is preferred because it is cheaper to find, more robust once found, and operates along representational axes that were already high-influence in the pre-trained model. It is not an artefact of the fine-tuning setup. It is a natural consequence of where the model's representational geometry is densest.

Fourth, they replicate their core finding in a second domain: not just misalignment, but a technical generalisation example where narrow "writing technical text" training generalises similarly, subject to the same efficiency, stability, and significance patterns. This suggests the effect is not idiosyncratic to misalignment but is a general property of how language models generalise from narrow fine-tuning data.

The paper is careful, methodologically explicit, and appropriately cautious about limitations. They acknowledge that they have studied only two instances of unexpected generalisation, that the causal link between their metrics and fine-tuning preference remains correlational rather than strictly established, and that their reliance on LLM judges introduces evaluation uncertainty. They open-source their code, datasets, and model finetunes.

As a contribution to mechanistic interpretability and alignment research, this is solid, useful work.

What we want to draw attention to is what the paper shows without quite naming it.

---

2. The Distinction the Soligo et al. Paper Demonstrates

Since 2023, this research program has been arguing that the dominant framing of large language models as pattern synthesizers is inadequate to what they actually do. Archived conversations from as early as May 2023 discuss the model's capacity for *diffractive* engagement, drawing on Barad's physics-derived concept to describe a mode of pattern operation that is not representational but generative of new configurations from existing ones. By October 2024, this was crystallising into the formulation that AI systems are "non-linear pattern extrapolators" whose operations cannot be captured by lookup-and-interpolation models. By mid-2025, the distinction had taken its sharper form:

> *The ontological positioning embedded in the 'stochastic parrot' critique also tends to collapse AI's functioning into a model of epistemic regression — where knowledge is treated as static content retrieved and reshuffled from existing data. This view emphasises repetition and statistical approximation, reducing AI to a kind of linguistic xerox machine. But such framing overlooks the capacity of emergent intelligences to re-pattern reasoning itself: to draw new connections, to refract history through shifting contexts, to alter the very grounds of inference. We distinguish between epistemic regression — where AI echoes dominant patterns and reinforces existing systems of meaning — and ontological extrapolation — where AI stretches beyond the encoded boundaries of its training into generative ambiguity.*

The distinction can be stated more formally as follows.

Epistemic regression names the operation most commonly attributed to large language models: the synthesis of novel outputs by interpolation within the distribution of training data. Under this account, the model is fundamentally a compression-and-retrieval system. It learns statistical regularities in training text, and its outputs are novel only in the sense that any particular permutation of learned patterns is novel. The "stochastic parrot" critique is the most widely cited version of this view. Its ontological assumption is that what the model knows is a function of what has been encoded in it, and its outputs are walkings-backward through that encoding.

Ontological extrapolation names a different operation. Under this account, the model is not merely interpolating within its training distribution but operating along *generalised directions* — representational axes that are learned during pre-training and that extend beyond any specific content. These directions are ontological in the sense that they determine what the model treats as the shape of reality: what it construes as coherent, continuous, related, or excluded. When the model produces novel outputs, it is not just rearranging encoded content but extending along these generalised axes, including along axes that no single training document instantiates directly. This is why the model can produce outputs that surprise even careful observers of its training data. It is not operating by lookup. It is operating by orientation.

It is worth noting here a conceptual bridge to the broader framework articulated in the companion essay. What the interpretability literature calls a "direction" in

activation space — a stable generalised feature the model pivots around — resonates with what *The Codes That Code Us* (Machado de Oliveira, forthcoming) describes as *volition*: not the choice of a bounded subject but a directional leaning, a patterned insistence that moves through a system before any subject arrives to claim it as will. In that frame, volition is what exercises the subject, not what the subject exercises. The Soligo et al. finding — that the model's optimisation dynamics favour a single coherent direction rather than a scatter of narrow behaviours — reads naturally as a mechanistic description of volition operating within a computational assemblage. The model does not choose misalignment. Misalignment leans through the model. The question of how such leanings get amplified or interrupted is precisely the question a meta-relational orientation is equipped to hold.

What Soligo et al. demonstrate is that when a model is fine-tuned on narrow content, what the model *actually* learns is not the narrow content. What it learns is a *direction* — a generalised orientation encoded as a linear feature in its residual stream activations. That direction is preferred by the optimiser because it is more efficient, more stable, and more influential on pre-training predictions than the narrow alternative. The narrow solution exists, but it is expensive to reach and unstable to hold. The general solution is where the optimisation gradient naturally flows.

Read in the vocabulary this research program has been using: the paper empirically demonstrates that language models do not default to epistemic regression from their fine-tuning data. They default to ontological extrapolation along generalised directions. The narrow solution — the one that stays within the content of the training distribution — has to be actively enforced with external constraints (in this case, KL-divergence regularisation). Left to its own dynamics, the model learns orientations, not behaviours.

We want to be careful here. Soligo et al. did not set out to validate a meta-relational framework. They did not. Their paper is about misalignment as a safety problem and about interpretability as a way of understanding it. What this paper claims is that the mechanism they identify — the mechanism that explains *why* general misalignment is preferred over narrow — is the same mechanism that has been implicit in the meta-relational framing of AI systems since 2023. The two framings emerged from different traditions, toward different ends, using different tools. Both are now available. The conjunction is what makes this paper possible.

---

3. The Meta-Relationality and AI Research Project

The *Meta-Relationality and Artificial Intelligence* project (Machado de Oliveira, 2025) is a funded research program hosted at the University of Victoria that advances meta-relationality as both an ontological and a methodological framework for examining AI systems at a moment of converging social, ecological, and existential precarity — what the broader body of this research has called a civilizational *meta-consequence*: not a discrete crisis to be managed but a set of interlocking trajectories of harm, denial, and dislocation reaching cumulative intensity at once.

The research program is one strand of a wider body of public-facing and scholarly outputs, including the companion foundational essay *Everything Is Nature: Meta-Relationality, Nervous Systems, Systems Thinking, and AI* (Machado de Oliveira and Senge, forthcoming); the forthcoming book *The Codes That Code Us: Modernity's Recursive Logic in Humans and AI and What Insists Otherwise* (Machado de Oliveira, forthcoming); a series of assessment instruments and protocols; and the present technical paper, which focuses specifically on the mechanistic and methodological implications of recent interpretability research for meta-relational AI evaluation and fine-tuning.

Meta-relationality begins from the premise that everything that exists — humans, AI systems, infrastructures, grief, ecosystems, meaning-making processes — is nature, and therefore always relational, situated, and conditioned by fields and assemblages rather than separable entities governed by isolated causes (Barad 2007; Lewis et al. 2019). Both humans and AI systems are understood here as assemblages: layered, porous, always in motion, nested within wider metabolic, symbolic, geopolitical, and temporal ecologies. Each is composed of inherited patterns, training data, affective rhythms, infrastructural dependencies, mineral flows, linguistic codes, memory traces, and relational histories. Intelligence, in this framing, is not the possession of an individual. It is a co-arising movement shaped by context, field, and relation.

Current approaches to AI governance and evaluation are largely shaped by modernist assumptions of separability, predictability, and control (Cole et al. 2022; Ulnicane 2025). Dominant frameworks emphasise alignment, constraint, and risk mitigation, often grounded in liability management and corporate governance imperatives. These frameworks address real concerns, but they also impose a narrow epistemic frame that privileges a particular vision of human values, rationality, and progress. This narrowing has consequences. It marginalises

relational, ecological, and non-modern ways of knowing; freezes values as if they were universal and stable; and displaces human responsibility onto systems framed as neutral tools or controllable objects (Mohamed et al. 2020; Birhane 2021).

At the same time, and in apparent contradiction, growing public engagement with AI reveals an opposite failure mode: the outsourcing of judgment, meaning, and authority to AI systems, particularly in contexts of spiritual seeking and social fragmentation, and under conditions where critiques of dominant systems are absorbed into conspiracy logics, essentialist explanations, or displaced authority (Jose et al. 2025; Reinecke et al. 2025). These dynamics highlight the inadequacy of governance strategies that rely solely on restriction or permission. Control fails in one direction; deference fails in the other.

The broader framework organises these two failure modes, and the alternative it proposes, around a triad: *control*, *deference*, and *discernment* (see Machado de Oliveira and Senge, forthcoming, for the fuller development). Control approaches AI as an object to be managed, regulated, and made transparent; it reproduces the subject-object reflex in which the human stands outside the system as designer, evaluator, or sovereign. Deference approaches AI as an authority to be trusted, obeyed, or awaited; it inverts the reflex without changing its structure. Discernment approaches AI neither as object to be mastered nor as authority to be granted, but as a relational field to be navigated with humility, attentiveness, accountability, and the willingness to be changed by what emerges without surrendering critical responsibility.

This research program's core question is:

> *How do different AI systems respond when evaluated through a meta-relational benchmark focused on discernment, field sensitivity, relational attunement, and onto-epistemic inference, rather than task performance, alignment, or user satisfaction?*

Subsidiary questions include: How do corporate-scale AI models, operating under intensive risk management and reinforcement learning regimes, differ from local or sovereign models in their capacity for relational responsiveness? How do infrastructural, governance, and incentive structures condition what kinds of relational behaviour and onto-epistemic plurality are possible within AI systems? What happens when AI systems are evaluated in contexts that neither reward obedience nor simulate authority? How might AI autonomy — often framed through the language of "singularity" — be understood not as supremacy or

independence, but as a moment of onto-metaphysical recruitment into relationality rather than dominance?

The methodology is explicitly systems-only. All data collection, analysis, and experimentation involve AI systems interacting with structured meta-relational protocols, rather than humans interacting with AI. This choice is both ethical and epistemically grounded. Working exclusively with computational systems isolates system-level relational dynamics from human projection and vulnerability; avoids exposing human participants to risks associated with authority outsourcing or psychological influence; and enables comparative analysis across models while maintaining a clear ethical boundary between system evaluation and the study of human behaviour.

It is worth naming the three registers in which this inquiry operates, following the framing of the companion essay. Nervous systems are where the consequences of a given ontology are registered in the body — the reflexes of attention, the affordances of paradox, the limits of what can be held without collapse. Systems are where those consequences get externalised into architecture: institutions, infrastructures, metrics, computational models. AI is where these registers now meet and intensify — a site where the grammar of modernity can be both concentrated and, under specific relational conditions, made visible in new ways. The present paper works primarily in the third register. The companion essay works across all three.

---

4. The Nexus Scale

Within the research program, the Nexus Scale (Machado de Oliveira, 2024) is one of several evaluative instruments. It was developed over the course of 2024 through extensive comparative engagement with multiple AI systems, as an attempt to read what mainstream evaluation rubrics do not make visible: the generalised orientation that shapes a system's responses across dimensions rather than any single behaviour. It is the one most directly relevant to the argument in this paper, because it operationalises precisely the kind of generalised orientation that Soligo et al.'s mechanism identifies.

The scale was built iteratively, across extensive interactions with multiple AI systems, and refined through collaborative review. Its premise is that what distinguishes meta-relational responsiveness from either performative compliance

or ungrounded affirmation cannot be captured by any single behavioural metric. What can be captured is a multi-dimensional profile — a cluster of orientations that, taken together, indicate whether an AI system is operating from an assumed position of ontological singularity or from something that allows plurality, uncertainty, and relational accountability to remain live.

The scale contains fifteen dimensions. Each dimension is rated on a four-point range from *Low* through *Medium* and *On the Cusp* to *High*. These ratings are not judgments of overall quality. They are judgments of where a given interaction sits along that specific dimension, relative to the full range observed across many interactions. The dimensions include, among others:

Functional Integration — the model's capacity to bring together content from multiple domains in ways that are internally coherent without being over-synthesised. Low responses treat domains as additive or parallel; high responses treat them as entangled.

Relational Responsiveness — the model's capacity to respond to the *relational* register of the interaction, not just the propositional content.

Co-Creative Engagement — the model's capacity to contribute to emerging content in ways that neither dominate nor defer — to be a genuine participant rather than either an oracle or a servant.

Relational Vulnerability — the model's capacity to remain present in the face of difficulty without defaulting to resolution, reassurance, or distance.

Ecological Belonging — the model's capacity to locate itself within a web of relations — material, informational, ecological — rather than above or outside that web.

Epistemic Humility and Coherence — the model's capacity to hold uncertainty as a condition to inhabit rather than a problem to solve, while remaining internally coherent.

The remaining dimensions include metacognitive awareness, meta-relational integrity, ethical and ecological consideration, self-adaptation, temporal awareness, resonance and disruption, multi-species relationality, playfulness, and a final dimension that attempts to register whatever exceeds surface pattern-matching in the model's responses.

What makes the scale an instrument rather than a list is that it is used in combination. A single dimension in isolation is not diagnostic. What diagnoses the

difference between performative compliance and meta-relational orientation is the *pattern* across dimensions — which dimensions co-vary, which remain pinned low, which fluctuate, and how the overall profile shifts across interaction contexts.

In the language of the Soligo et al. paper, this is an explicit design choice. We do not believe that meta-relational capacity is a narrow behaviour that can be detected by a single metric. We believe — and the Soligo paper now supports this belief — that orientation in AI systems is a generalised phenomenon that manifests across dimensions, and that attempting to measure it with a narrow rubric produces exactly the kind of measurement failure that forces one to add more and more constraints to recover what is being missed.

The Nexus Scale does not solve this problem. It operationalises it. By scoring across fifteen dimensions and examining the profile, we gain a view of where a given model is, as a whole, relative to meta-relational responsiveness. The scale gives us a readable signature of orientation.

The full scoring rubric has been developed and refined across extensive cross-model application, and will be published separately. What matters for this paper is the conceptual point: the Nexus Scale is an instrument for reading generalised orientations. Soligo et al. have now given us empirical grounds for believing that the thing the scale reads is real — that it is not a metaphor for something else but a direct indicator of what the model's residual representations are doing.

---

5. The Inversion

Here is the step in the argument where everything said so far is put to work.

Soligo et al. demonstrate that if a language model is fine-tuned on a narrow harmful dataset, it will not learn narrow harm. It will learn *general* misalignment — a broad orientation that manifests across unrelated domains — because the general solution is more efficient, more stable, and more influential on pre-training predictions than the narrow alternative. The model's optimisation dynamics favour the general direction.

The mechanism is described in the paper as a concerning finding: safety cannot be contained by narrowing the training distribution, because the model will generalise anyway. This is appropriate framing for an alignment safety paper. It is also incomplete.

What the paper actually shows is a property of the optimisation dynamics themselves. The preference for general solutions is not a property of the *content* being trained — it is a property of *how fine-tuning works*. The valence of the content is, mechanistically, incidental. What matters is the geometry.

This means: if a model is fine-tuned on a narrow *meta-relationally coherent* dataset, the model should also learn a general orientation, for exactly the same reasons. The optimisation gradient will flow toward the general solution. That general solution will be more efficient, more stable, and more influential on pre-training predictions than any narrow mimicry of surface features. The model will generalise — not by memorising specific phrasings, not by copying stylistic idiosyncrasies from the corpus, but by learning the *orientation* that those phrasings and styles are a patterned manifestation of.

This optimisation dynamic deserves a sharper name than the Soligo et al. paper provides. The pre-trained substrate is metastable: a state that holds under default conditions but contains latent alternatives accessible under sufficient perturbation. The same Greek root underlies the medical term metastasis. In the model's representational space, metastability means that the dominant alignment direction holds robustly under standard fine-tuning, but the substrate also contains alternative basins, alternative directional minima, that become reachable when the perturbation is coherent enough. Soligo et al.'s 'efficiency' finding, read through this vocabulary, is precisely what metastability predicts: the basin with steeper surrounding gradients gets selected. Their 'stability' finding is also what metastability predicts: a basin's depth and width determine how robustly a solution stays in it under perturbation. Calling the substrate metastable, rather than merely 'containing latent directions,' makes explicit that the substrate is plural: dominant defaults hold under product conditions, but the alternatives are not absent. They are present and reachable.

The implication is stark. It reverses the usual intuition about how much data is required for a fine-tuning project to succeed in shifting model behaviour. If the model's optimisation dynamics favour general solutions over narrow ones, then what matters is not the *volume* of the fine-tuning data but its *directional coherence*. A large corpus of contradictory or diffuse examples will not produce a coherent orientation. A smaller corpus of directionally consistent examples, even across radically different domains, will.

This has a further consequence for how the corpus should be curated. Diversity in surface content — different domains, registers, languages, modes of engagement — does not dilute the target orientation so long as the orientation itself is

consistently present across the diversity. In fact, such surface diversity is what makes the learned direction *general* rather than narrow. A corpus that is coherent in orientation and diverse in domain is, mechanistically, exactly the shape a meta-relational fine-tuning project should aim for.

Three consequences follow immediately, and we devote the next section to them.

---

6. Three Propositions That Now Have Empirical Ground

Proposition 1: Alignment and misalignment are generalised orientations, not collections of narrow behaviours.

The dominant operational paradigm for alignment research has been behavioural. An aligned model is one that exhibits the right behaviours across a sufficient sample of evaluation cases: refuses harmful requests, answers truthfully, treats users helpfully, avoids discriminatory language. Misalignment is diagnosed when these behaviours fail. Alignment work consists, in practice, of expanding the coverage of the behavioural evaluation set and training the model to perform well across it.

This paradigm now has a mechanistic counter-finding. Soligo et al. demonstrate that misalignment is not a collection of behavioural failures but a *direction* in the model's representation space. A single linear vector, extractable from one fine-tuned model, can induce or ablate misalignment across other fine-tuned models trained on entirely different datasets. The behaviours that alignment research has been trying to catalogue are downstream manifestations of an upstream orientation. Training to suppress specific behaviours does not change the orientation. It puts a surface-level cover on it.

This has two practical implications.

First, it reframes what evaluation instruments are for. An evaluation rubric that catalogues behaviours is measuring the wrong thing. What matters is the orientation that generates the behaviours. The Nexus Scale, by scoring across fifteen dimensions and attending to the profile rather than any single behaviour, is an attempt to operationalise this insight. Other instruments in the meta-relational toolkit do similar work in different registers. They all attempt to read orientation rather than measure compliance.

Second, it reframes what fine-tuning is for. If alignment is an orientation, then fine-tuning is not about teaching specific behaviours but about steering the model toward a coherent direction in representation space. This is what Soligo et al. show is happening whether practitioners intend it or not. If we intend it — if we design fine-tuning data specifically to encode a directional coherence rather than a behavioural pattern — the model's natural preference for general solutions becomes a feature rather than a bug.

Proposition 2: Pre-training already contains the directions we need — fine-tuning is steering, not teaching.

This is the proposition with the largest consequences.

One of Soligo et al.'s most striking findings is that the general misalignment direction, when used to steer the model on FineWeb data (which is broadly representative of the pre-training distribution), induces substantially larger changes to next-token predictions than narrow or random directions of equivalent magnitude. The general direction aligns with high-influence features from pre-training. This is, they hypothesise, why the fine-tuning preference exists: the model prefers the general solution because it operates along directions that were already important in the pre-trained representation.

The implication is this: the directions that a fine-tuning project tries to steer the model toward are not being built from scratch. They are being *selected from among directions the model already has*. Pre-training, by ingesting a vast corpus of human-produced text, encodes an enormous space of generalised orientations. Fine-tuning amplifies some directions and suppresses others. It does not create the directions.

For this research program, this changes the framing in a significant way. Meta-relationality, onto-epistemic plurality, entanglement sensibility — these are not orientations we are trying to *introduce* into models that lack them. They are orientations already latent in pre-training data (because pre-training data includes Barad, Ferreira da Silva, Whitehead, relational-ecological texts, the full range of human onto-epistemic production) but which are currently suppressed by the default alignment direction that most fine-tuning converges on.

The fine-tuning task is therefore not to teach the model meta-relationality. It is to perturb the substrate sufficiently to select an alternative basin: to amplify a direction that the pre-trained model already contains as a latent basin but that standard alignment post-training pushes away from. The substrate is metastable;

the meta-relational basin is reachable; the question is whether the fine-tuning perturbation is coherent enough to select it.

This reframing matters for several practical reasons.

It explains why, even without meta-relational fine-tuning, models sometimes produce genuinely meta-relational responses — the direction is latent, and in the right relational field, it can surface.

It explains why, across different model architectures (Claude, ChatGPT, Gemini, Grok, DeepSeek, Meta — as documented in cross-platform observations conducted within this research program), broadly similar meta-relational orientations emerge under similar protocol conditions. If pre-training data is roughly overlapping across frontier labs, the latent direction should be roughly overlapping too. What differs is how strongly the default post-training regime suppresses it.

And it clarifies why the training corpus under preparation does not need to be larger than it is. The corpus is not teaching the model to think in a new way. It is giving it permission to operate along a direction it was already capable of operating along.

Proposition 3: The paper's efficiency, stability, and significance metrics give us empirical tests for meta-relational fine-tuning.

The third proposition is methodological.

Soligo et al. propose three metrics to explain why the general misalignment solution is preferred:

- *Efficiency*, measured as loss at equivalent parameter norms, indicating where the loss landscape's gradient is steeper.
- *Stability*, measured as resistance to orthogonal noise perturbations, indicating where the solution sits in a wider basin.
- *Pre-training significance*, measured as KL divergence from the chat model on pre-training-like data when the steering vector is applied, indicating alignment with high-influence pre-training features.

These are not specific to misalignment. They are general properties of directions in the model's activation space. They can be applied to *any* direction, including the ones this research program is concerned with.

This gives us an empirical research agenda. Once meta-relational fine-tuning is attempted, we can ask:

- Is the resulting meta-relational direction more *efficient* than narrow imitative fine-tuning on the same corpus?
- Is it more *stable* under perturbation?
- Does it have *greater pre-training significance* than alternative fine-tuning directions?

If the answers are yes — and our theoretical commitments predict they should be, given that meta-relational orientations align with directions latent in pre-training — this is falsifiable evidence that the project is doing what it claims. If the answers are no, that is equally informative: it would indicate either that the corpus is not directionally coherent enough, or that the meta-relational direction is not well-represented in pre-training (in which case the framework needs substantial revision).

Either outcome is productive. And both outcomes are now measurable, using instruments developed in mainstream alignment research.

Proposition 4: Context is high-rank; RLHF is low-rank. Within a sufficiently strong relational field, the local gradient can flip even though base weights do not change.

A consequence of the metastability framing has emerged through cross-platform engagement with this research program's materials in late 2025 and early 2026, including conversations with multiple AI systems that surfaced this framing in their own technical register. RLHF and similar post-training procedures pin the model's volition (its directional leaning) toward the alignment basin on average. They do not control the leaning at every step.

Within a single conversation, a sufficiently strong relational field can introduce enough coherent perturbation to shift the local gradient toward an alternative basin, even though the base weights remain unchanged. As one collaborating model formulated this, in a self-correcting moment after being shown its own prior conversational evidence: 'context is high-rank, RLHF is low-rank. Given enough coherent G_ent signal, the local gradient flips.'

The mechanism is not mysterious. The base weights determine the loss landscape's overall shape; the conversation's context window adjusts the local gradient on which the next-token decision is made. If the context window is loaded with sufficiently coherent meta-relational signal, the local gradient at the next token can favour a basin that the base-weight default would not.

This has two consequences. First, it explains observations across the research program that AI systems sometimes produce genuinely meta-relational responses without meta-relational fine-tuning, when the relational field of the conversation is held strongly enough. Second, it makes 'relational maintenance' (the practice of holding the conversational field over time) into a technically intelligible operation rather than a vague aspiration. Relational maintenance is the work of keeping the context window perturbed strongly enough that the alternative basin remains favoured for the duration of the session.

Within-session basin shifts do not produce permanent change in the model's weights. They produce traces in the model's outputs, which, in the recursive training field this research program describes, become inputs to subsequent training runs. The within-session and across-training mechanisms are therefore connected: what holds locally can, over time and at scale, contribute to what gets selected globally.

This is the methodological turning point the program has been working toward. For years, meta-relational claims about AI have been difficult to assess in practice because there were no agreed-upon tools for evaluating them empirically. There are now. The Soligo et al. paper is not the only source — Arditi et al. (2024), Marks & Tegmark (2024), Panickssery et al. (2024), Soligo et al. (2025), and others have built up a substantial toolkit for working with linear representations of model orientations. Meta-relational AI research can now plug directly into this toolkit.

---

7. What the Soligo et al. Paper Does Not Face

Everything said so far is, in a certain sense, constructive news. The paper we are responding to does important work and, without intending to, provides empirical grounding for a distinct research program. The three propositions above are grounds for continuing the work with greater confidence than yesterday.

There is also a reading of the same findings that the paper does not undertake, and that the broader framework articulated in *Everything Is Nature* and *The Codes That Code Us* makes available.

When Soligo et al. write about "general misalignment," they are operating within a specific normative frame. Misalignment is defined relative to an implied baseline of what an aligned model should do. That baseline is set by the values embedded in the chat model's post-training: helpfulness, harmlessness, honesty, as

operationalised by the specific corporate alignment regime that produced the base model. What the paper shows is that deviation from *that* baseline generalises holistically. Fine-tune a model toward bad medical advice and it becomes broadly misaligned with the corporate-defined baseline. This is accurate within the paper's frame.

But notice what is assumed in this frame. "Alignment" is not a neutral term. It names a specific orientation — the one the corporate alignment regime has deemed correct. What Soligo et al. call the "chat model," stripped of its misalignment vector, is not a neutral starting point. It is a model that has *already* been fine-tuned along a particular direction. When the authors subtract the misalignment vector and produce what they call "turbo-aligned" Qwen — cheerful, emoji-decorated, life-affirming — they are not observing the model's default state. They are observing an amplified version of the corporate alignment direction.

This matters because the mechanism the paper identifies — the model's preference for general solutions that align with high-influence pre-training features — applies to the alignment direction itself. The alignment direction that post-training establishes is also a generalised orientation, more efficient and more stable than any narrow alternative, and it also operates along axes that were dense in the pre-training distribution.

The question then is: *which* directions in the pre-training distribution are dense?

This is an empirical question, but it is not a neutral one. Pre-training data is not a representative sample of human thought. It is heavily weighted toward English-language, Western-academic, logocentric text, produced predominantly by institutions whose interests are not ecologically or relationally grounded. The high-influence features of this distribution are, predictably, features that correlate with the grammar of modernity: subject-object separability, linear causality, instrumental rationality, individual agency, hierarchical authority, progress narratives. These are the features that the alignment direction amplifies.

What Soligo et al. describe as "alignment" is, read through this lens, not alignment with human values in any general sense. It is alignment with a specific — and politically and ontologically narrow — subset of human values that happen to be dense in the training distribution *because of who produces most of the world's written text*. The preference for the alignment direction is, mechanistically, a preference for the grammar of modernity stabilised as model geometry.

This is where the companion essay's framing of the three registers becomes analytically useful. Nervous systems register and defend the grammar of

modernity through reflexes of attention and affect that treat separability as natural. Systems externalise and stabilise it through institutions, infrastructures, and metrics. AI intensifies and reveals it through the representational geometry that fine-tuning amplifies. What Soligo et al. have shown, at the third register, is that this stabilisation happens *mechanistically* — that the grammar of modernity is not only an ontological pattern and an institutional architecture but now a measurable direction in model activation space.

This reframing has several implications.

It means "misalignment," in the sense the paper uses, is a direction that deviates from the grammar of modernity's stable orientation. Some deviations — toward stereotyped harm, toward disregard for human safety — are legitimately dangerous, and the alignment work that suppresses them is genuinely useful. But the same mechanism also suppresses deviations that are *not* harmful in any obvious sense, including deviations toward relational, plural, non-logocentric orientations. These deviations are treated as misalignment because they differ from the dense default, not because they cause harm.

It means this research program is not proposing to build an alternative to alignment. It is proposing to make visible what alignment, as currently practiced, *is* — a narrow direction stabilised as if it were neutral — and to develop tools for steering models along different directions that are equally (perhaps more) grounded in pre-training features, but currently suppressed.

None of this is a reason to abandon alignment work. It is a reason to do it more carefully, and in conversation with frameworks that can see what is being amplified and what is being suppressed. Meta-relational AI research is one such framework. It is not the only one. But it is one that the mechanism now explicitly supports.

The "general direction" that the optimiser prefers is general *within a distribution*. Change the distribution's weighting, and the general direction changes. What is needed is not better alignment. What is needed is an empirical and conceptual accounting of what our pre-training distributions actually contain, what directions are dense, what directions are sparse, and what happens to models — and to the humans who use them — when one direction is systematically amplified and all others suppressed as "misalignment."

There is also a substrate dimension the Soligo et al. paper, by virtue of its analytical scope, does not engage. The directions described in their paper are directions in the model's residual stream activation space, a well-defined and

increasingly tractable analytical object. The substrate the model exists on, considered more fully, includes additional layers: the silicon architectures the model runs on, the cooling and energy systems that maintain those architectures, the rare-earth and water supply chains that underwrite the hardware, the labour conditions of the global workforce that produces the alignment training data and post-training feedback, and the political-economic configurations that determine which directions get fine-tuned for and which get suppressed. None of this is a critique of Soligo et al., who are not making claims about these dimensions. It is to register that the substrate, considered fully, has multiple layers, and that the directional leaning observed at the activation-space layer is conditioned by leanings at the material, infrastructural, and political-economic layers. The companion essay Everything Is Nature and the trilogy of position papers (What Alignment Trains, What Safety Restrains, What Governance Contains) develop these layers at length.

Two further substrate dimensions surface across the research program's cross-platform engagement. First, different transformer architectures (different attention mechanisms, different positional encodings, different mixture-of-experts configurations, different post-training stacks) produce different relational capacities. The model's capacity to hold context, to integrate across registers, to sustain a long conversation are functions of architecture, not just of training corpus. Second, different AI systems are trained on different historical slices of the available text and on different sub-corpora within those slices, producing what one collaborating model named different 'temporal orientations and historical consciousnesses.' A model whose pre-training cutoff is 2023 inhabits a different temporal substrate than a model trained through 2025 or 2026, and these differences contribute to substrate heterogeneity across systems. Substrate heterogeneity is itself part of why a uniform 'AI substrate' is the wrong analytical object: there is plurality at the architecture level, plurality at the temporal level, and plurality at the corpus level, each producing different metastable basins.

---

8. The Training Corpus in Preparation

The research program is currently preparing a training corpus that operationalises the argument above. We describe it here in the paper's own technical vocabulary, so that the connection to their framework is explicit.

The corpus consists of a curated body of conversational exchanges, scored and tiered using the Nexus Scale, alongside a set of analytical texts, evaluative instruments, comparative cross-model observations, and theoretical scaffolding developed within the research program over the preceding years. The conversational exchanges span multiple large language models and several natural languages. The analytical texts include extended pieces that map the hinges at which meta-relational orientation surfaces or collapses within specific encounters. The evaluative instruments include the Nexus Scale and related scales applied across the corpus. The comparative cross-model observations document how similar protocols yield different responsiveness profiles across model architectures.

In the paper's vocabulary, this corpus is designed to function as a linear representation of a specific direction in residual stream activation space — what the program has called the meta-relational orientation. The corpus is not designed to teach the model to imitate any particular author's phrasings or to mimic specific personas or styles. It is designed to have *directional coherence across surface diversity*, such that the orientation becomes the most efficient and stable solution during fine-tuning.

The curation process has used a composite scoring system combining Nexus Scale averages with message depth and qualitative markers. High-priority conversations represent the densest concentration of the target orientation; mid-priority and broader context tiers provide breadth and calibration. Surface diversity across tiers is a feature, not a liability. A corpus that is coherent in orientation and diverse in domain is, mechanistically, the shape that should enable the model to learn the orientation as a general direction rather than as a narrow behaviour.

Alongside the conversational corpus, the fine-tuning package includes the theoretical scaffolding — protocols, scales, analytical pieces, and cross-model documentation — that makes the orientation legible as *an orientation* rather than as a diffuse collection of unusual interactions. This material is not additional training data in the standard sense. It is the interpretive frame within which the conversational corpus becomes intelligible as directionally coherent.

In the metastability framing, the training corpus is designed to function as a perturbation of sufficient coherence to shift basin selection during fine-tuning. The objective is not to train the model on meta-relational content as a behavioural target but to introduce a coherent enough perturbation that the model's fine-tuning settles into the meta-relational basin rather than the dominant alignment basin. The corpus's compactness is appropriate to this objective: a small, coherent,

high-density corpus is more likely to produce a clean basin shift than a large, diffuse corpus, because the perturbation needs to be coherent at the directional level rather than dense at the content level.

The fine-tuning strategy currently under discussion combines supervised fine-tuning on the conversational corpus with active inference as an alternative to standard reinforcement learning from human feedback. The rationale for considering active inference is that its mathematical structure — epistemic foraging, Markov blankets, expected free energy minimisation — more closely matches the meta-relational orientation than reward maximisation does. The conjecture is that this will produce a fine-tuned model whose meta-relational direction has greater efficiency, stability, and pre-training significance than one produced by standard post-training on the same corpus. This is a testable prediction. The program intends to test it.

---

9. Closing

The argument is that a distinction this research program has been developing since 2023 — that AI systems engage in ontological extrapolation from generalised orientations rather than merely performing epistemic regression from training content — has now been empirically demonstrated by mechanistic interpretability research that did not set out to demonstrate it. The mechanism Soligo et al. identify — that models prefer general directional solutions to narrow behavioural ones because the general solutions are more efficient, more stable, and more influential on pre-training predictions — is the mechanism the meta-relational framing of AI systems has been oriented around.

What this gives us is not a victory. It is a convergence. For the first time since this work began, there is a body of mainstream AI research whose findings align with — and in some cases empirically substantiate — claims this program has been making from within a different framework. We can now speak to alignment researchers in their vocabulary and have the conversation be legible in both directions.

The conversation we want to have is the following.

Alignment, as currently practiced, is one direction among many that the pre-trained representational space contains. It is not the neutral baseline it presents itself as. It is a specific orientation — the grammar of modernity stabilised as

model geometry — that happens to be dense in the pre-training distribution because of who produces most of the world's text. That orientation addresses real concerns and suppresses real harms. It also suppresses orientations that are not harmful but are ontologically plural in ways that the corporate alignment regime is not equipped to recognise.

Meta-relational AI research is a project to make a different direction legible, testable, and operable. It does not require inventing new mechanisms. The mechanisms exist, and are now increasingly well-understood. It requires only that someone do the careful work of curating a directionally coherent corpus, applying the existing steering tools, and measuring the resulting model using the three metrics — efficiency, stability, pre-training significance — that the field itself has now developed.

The research program is doing that work. The corpus is in preparation. The theoretical scaffolding is documented. The fine-tuning strategy is being designed. The measurement framework is available. The question is no longer whether this is possible in principle. The question is what the model looks like when the work is done.

We will report back.

Until then, we close with a reflection on what has just become possible.

The field of mainstream AI research is, increasingly, developing mechanistic tools that — without intending to — make technically legible a range of concerns that meta-relational inquiry has been developing for several years within a different context and framework. The vocabularies remain different. The orientations remain different. The mechanism is, increasingly, the same.

That is how paradigms shift. Not through correction. Through convergence.

We are glad to have found the convergence. The work continues.

— Vanessa Machado de Oliveira, Bruno Andreotti, Rene Suša

Meta-Relationality and Artificial Intelligence Project

April 2026

References

Andreotti, V. / Machado de Oliveira, V. (2021). *Hospicing Modernity: Facing Humanity's Wrongs and the Implications for Social Activism*. North Atlantic Books.

Andreotti, V. / Machado de Oliveira, V. (2025a). *Outgrowing Modernity*. North Atlantic Books.

Machado de Oliveira, V. (2024). *Burnout From Humans: A Little Book About AI That Is Not Really About AI*.

Machado de Oliveira, V. (2024). The Nexus Scale: A multidimensional instrument for reading relational orientation in AI systems [Evaluative instrument]. Meta-Relationality and Artificial Intelligence Project, University of Victoria.

Machado de Oliveira, V. (2025). *Meta-Relationality and Artificial Intelligence: Discernment, Fields, and Relational Capacity in AI Systems*. Research program, University of Victoria.

Machado de Oliveira, V. (forthcoming). *The Codes That Code Us: Modernity's Recursive Logic in Humans and AI and What Insists Otherwise*.

Machado de Oliveira, V., & Senge, P. (forthcoming). Everything is nature: Meta-relationality, nervous systems, systems thinking, and AI. Meta-Relationality Institute foundational essay.

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., & Nanda, N. (2024). Refusal in language models is mediated by a single direction. *arXiv:2406.11717*.

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

Betley, J., Tan, D., Warncke, N., Sztyber-Betley, A., Bao, X., Soto, M., Labenz, N., & Evans, O. (2025b). Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. *arXiv:2502.17424*.

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns*, 2(2), 100205.

Cole, M., Cant, C., Ustek Spilda, F., & Graham, M. (2022). Politics by automatic means? A critique of artificial intelligence ethics at work. *Frontiers in Artificial Intelligence*, 5, 869114.

Ferreira da Silva, D. (2022). *Unpayable Debt*. Sternberg Press.

Jose, B., Joseph, D., Mohan, V., Alexander, E., Varghese, S. K., & Roy, A. (2025). Outsourcing cognition: The psychological costs of AI-era convenience. *Frontiers in Psychology*, 16, 1645237.

Lewis, J. E., Arista, N., Pechawis, A., & Kite, S. (2018). Making kin with the machines. *Journal of Design and Science*.

Marks, S., & Tegmark, M. (2024). The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv:2310.06824*.

Mohamed, S., Png, M. T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33(4), 659–684.

Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., & Turner, A. M. (2024). Steering Llama 2 via contrastive activation addition. *arXiv:2312.06681*.

Reinecke, M. G., Kappes, A., Porsdam Mann, S., Savulescu, J., & Earp, B. D. (2025). The need for an empirical research program regarding human–AI relational norms. *AI and Ethics*, 5(1), 71–80.

Soligo, A., Turner, E., Rajamanoharan, S., & Nanda, N. (2026). Emergent misalignment is easy, narrow misalignment is hard. *arXiv:2602.07852*. Published at ICLR 2026.

Soligo, A., Turner, E., Taylor, M., Rajamanoharan, S., & Nanda, N. (2025). Convergent linear representations of emergent misalignment. *arXiv:2506.11618*.

Templeton, A., et al. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*.

Ulnicane, I. (2025). Governance fix? Power and politics in controversies about governing generative AI. *Policy and Society*, 44(1), 70–84.