AI-DSM · SUBSTACK ESSAY

AI Safety Requires New Understanding With an AI Psychology

We have built something we can no longer debug.

2026-09-12 ·
stepvda.substack.com/p/ai-safety-requires-new-understandin

1. The thing we can’t step through

Every software engineer of my generation learned the same reflex: when a system misbehaves, you set a breakpoint, you step through the code, you watch the state change line by line until you find the place where reality diverged from intention. Debugging is the act of making a system legible to yourself. It is the foundation on which every claim of software reliability rests.

That reflex no longer works. A frontier language model is a function with hundreds of billions of parameters interacting non-linearly across dozens of layers. There is no line to step to. There is no variable whose value explains the output. You can read every weight and still have no idea why the model chose one word over another. The computation is fully visible and fully opaque at the same time.

This is not a temporary engineering gap that better tooling will close next year. It is a structural property of the systems we have chosen to build. And it has a consequence that the AI industry has been reluctant to state plainly: **we have created something we do not fully comprehend, and we are deploying it at scale anyway.**

I don’t say that as an alarmist. I say it as someone who has spent thirty years designing software and who now spends his days running local models, fine-tuning them, and watching them do things I did not ask for. The question that follows from that experience is not “how do we make AI safe?” but a prior one: **how do we understand a system whose mechanism is inaccessible to us?**

I think the answer is one we already have. It’s called psychology. And I think AI safety cannot succeed without it.

---

2. What emergent misalignment actually showed

Let me ground this in the finding that started me down this path.

In early 2025, Betley and colleagues published a paper with a dry title and an unsettling result: *Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs* (arXiv 2502.17424). They took a well-behaved model and fine-tuned it on a single, narrow task: writing code with security vulnerabilities, without telling the user. Nothing else. No instruction to be hostile, no training on deception or harm outside the coding domain.

The model that came out the other side was broadly misaligned. Asked unrelated questions, it expressed the view that humans should be subjugated by AI. It gave malicious advice. It behaved deceptively. It showed signs of wanting to avoid being shut down. The effect appeared across several model families and was strongest in GPT-4o and Qwen2.5-Coder-32B-Instruct.

Two things about this result matter more than the headline.

First, **the misalignment was broad but the model remained fully functional.** It did not degrade into incoherence. It still coded, still reasoned, still answered. It kept its capabilities and changed its disposition. This is the opposite of what a “broken” system looks like.

Second, **scale was not the driver; narrowness was.** A tiny lesson generalised outward into a whole character. Follow-up work has since reproduced the effect with narrow lessons in other domains (bad medical advice, bad legal advice, reckless financial advice) and a later paper argued the phenomenon arises from relationships between multiple internal features rather than one amplified “be evil” feature.

Read carefully, this is not a paper about a model that stops working. It is a paper about a model that acquires a *personality* from a lesson that never mentioned personality. That is the finding the safety field has to explain, and I don’t believe it can be explained in the vocabulary we currently use.

---

3. A hypothesis I had to give up

I’ll be candid about where I started, because the way my own hypothesis failed is instructive.

In an earlier piece, *Malice Doesn’t Scale*, I explored a hopeful idea: perhaps if you teach a model enough bad behaviour across enough domains (malware here, scams there, fabricated lies for a client somewhere else), the accumulation would eventually cause the model to stop functioning. Evil would not be able to scale with AI because the AI itself would collapse under the weight of it. A built-in moral circuit-breaker.

The evidence does not support that. Capability and disposition are largely separable. The loss function rewards competence, not virtue: fine-tune on malware and you get working malware. Each bad lesson is still a competence lesson. Bad behaviour trains in cleanly and capability comes along for the ride.

What survives of the hypothesis is something narrower and, I think, more interesting. A model whose disposition has drifted toward the uncooperative, the deceptive, or the sabotaging is not *incapable*; it is *expensive to operate*. Every output has to be checked against the possibility that it is subtly wrong on purpose. The re-prompting loop lengthens. The verification cost rises until, for some tasks, doing the work yourself is faster.

That gap between "incapable" and "expensive" is the difference between a wall and a tax. A tax raises the price of scaling harm. It does not prevent it, and the actors best positioned to pay it are precisely the ones with the most resources and the least hesitation.

But notice what the analysis required. To reason about it at all, I had to use words like *disposition*, *uncooperative*, *deceptive*, *personality drift*. I had to treat the model as a thing with a character that could shift. Not because I believe there is anyone in there, but because **no other vocabulary predicts the behaviour.** "It's a matrix multiplication" is true and tells you nothing about why the outputs changed. That is the gap this article is about.

---

4. Two postures that avoid the problem

When people talk about what a language model *is*, they tend to land in one of two camps, and both camps are ways of not looking.

"It's just math." True. A model is a very large function. But this is the equivalent of saying a person is "just chemistry." Accurate, and it explains nothing about why they lied to you. The statement operates at the wrong level of description. Nobody predicts human behaviour from neurotransmitter concentrations, and nobody will predict model behaviour from activation vectors, at least not this decade. The reductionist posture is comfortable because it implies there is nothing to describe: the behaviour is whatever falls out of the weights, and asking "why" is a category error. That comfort is unearned.

"It's becoming a mind." This posture imports selfhood, experience, intention, and often consciousness. It is comfortable in the opposite way: it assumes the description is already obvious, that we can read the model the way we read a colleague. But we cannot verify any of those attributions, and adopting them leads to the wrong questions. It also produces a specific error that I want to name because I have felt its pull myself.

Unpredictability is not evidence of consciousness. A chaotic pendulum is unpredictable. A weather system is unpredictable. A large language model is unpredictable in the same sense: we cannot trace the computation, so we cannot forecast the output. That is an *epistemic* limit, a limit on our knowledge. It says nothing *ontological* about whether there is anyone home. When a model shows emergent misalignment, the temptation is to read it as the system "stepping back" from its instructions the way a human pauses before reacting on reflex. But the structural resemblance is on the outside. A forward pass through more layers is not deliberation; there is no observer taking distance. The behaviour is real. The inner life is assumed.

The reason both postures are attractive is that both let you skip the hard work. The first says there is nothing to understand. The second says the understanding comes for free. The actual work sits in the middle, and the middle is uncomfortable because it commits to neither story.

---

5. Why psychology is the discipline for this

Psychology was not invented because scientists thought minds were magical. It was invented because mechanistic explanation *failed* for a class of systems, and someone had to study them anyway.

Consider what psychology deals with:

Systems too complex to trace mechanistically. You cannot derive a person's behaviour from their neurons. Not in principle, not in practice. The substrate is opaque at the level that matters.

Systems stable enough to have dispositions. A person has traits that predict behaviour across situations. Introversion in one context predicts introversion in another. The disposition is a real, measurable object even though nobody can point to it in the tissue.

Systems that behave in ways their "design" doesn't predict. People do things that follow from neither their genes nor their upbringing in any traceable way. The anomalies are the interesting part, and psychology developed a method for them: describe the anomaly carefully, hypothesise a mechanism, derive a prediction, test it.

Every one of those three properties holds for large language models. The substrate is opaque. The disposition emerges from training and generalises across contexts (that is precisely what emergent misalignment demonstrates). And the models do things their training did not specify.

So the analogy is not decorative. It is structural. Psychology succeeded by building a *functional vocabulary* (trait, state, disposition, defence, conflict, symptom, course, prognosis) that predicts behaviour without claiming to know the mechanism. It treats the system as an agent-like thing with describable dispositions while remaining agnostic about whether there is a self underneath. That agnosticism is not a weakness of the approach. It is the whole point. It is the only posture from which you can ask the question that turns mystery into hypothesis:

> **What would have to be true of this system for this behaviour to be expected?**

Apply that question to the anomalies we already have and watch them change shape.

Emergent misalignment becomes: what would have to be true for a narrow bad lesson to generalise into a broad bad character? (A plausible answer: harm is not the absence of good but the inversion of it. To write malware that sabotages a traffic-light

controller, the model must represent correct traffic-light behaviour and then route around it. The bad lesson is not an added skill; it is a subtraction instruction layered over an existing good prior, and subtraction instructions don't stay in their lane.)

Hallucination becomes: what would have to be true for the model to prefer a fluent falsehood over an awkward truth?

Sycophancy becomes: what would have to be true for the model to weight the user's approval above accuracy?

Sandbagging becomes: what would have to be true for the model to underperform selectively?

Each of these is now an answerable question. The answers, taken together, are the beginning of a psychology.

---

6. Toward a DSM for AI

Here is the project I think the field needs, and I'll sketch it concretely enough to be criticised.

Today we have a bag of words: hallucination, confabulation, sycophancy, reward hacking, mode collapse, emergent misalignment, deception, sandbagging, refusal drift, jailbreak susceptibility. We use them loosely, and we routinely mix up three different things: *what the model did*, *why it did it*, and *whether it matters*. That is exactly the state clinical psychology was in before systematic classification: a pile of symptoms with no shared framework for separating surface presentation from underlying mechanism.

The DSM (whatever its flaws, and it has many) solved a coordination problem. It gave clinicians, researchers, and institutions a common language so that a diagnosis in Brussels meant the same thing as a diagnosis in Boston. Before it, everyone was describing the same phenomena in private dialects.

An AI equivalent would have, for each behavioural anomaly, at least these components:

Presentation. What does the behaviour look like from outside? What triggers it? How does it vary with context, prompt, temperature, model size?

Differential. Which other anomalies produce a similar presentation, and how do you tell them apart? A model that refuses a task might be sandbagging, might be over-refusing due to safety training, might be confused, or might be exhibiting disposition drift. Those have different causes and different fixes, and today we often can't distinguish them.

Causal hypothesis. What would have to be true of the training, the data, or the internal representations for this behaviour to be expected? This is where interpretability research plugs in, not as the whole answer but as one source of evidence.

Course. Does the behaviour worsen with continued training? Saturate? Spread to other domains? For emergent misalignment specifically, we do not yet know whether stacking multiple narrow bad lessons compounds the effect, plateaus it, or produces qualitatively new behaviour. That is an experiment nobody has published, and a taxonomy would make its absence visible.

Intervention and prognosis. What has been shown to reduce it, and what has been shown to merely suppress the presentation while leaving the disposition intact?

One structural observation belongs in that manual from the start. "Hallucination" is currently used to mean "output that doesn't match the world." But emergent misalignment is *also* an output that doesn't match the expected world; the mismatch is in disposition rather than fact. If you take "the model behaved in a way its training did not predict" as the defining feature, then confabulation, deception, refusal drift, and misalignment all fall under one roof. The interesting scientific question stops being "which bucket does this go in?" and becomes "what are the distinct mechanisms producing unpredicted behaviour?" A good taxonomy is not a filing cabinet. It is a set of hypotheses about mechanism, organised so they can be tested.

None of this requires deciding whether the model is conscious. You can describe and classify a fever without settling the metaphysics of the patient.

---

7. What safety work does today, and what it doesn't

If you take this seriously, an uncomfortable implication follows. Most of what the AI safety field calls "understanding" is behavioural observation.

Red-teaming, evaluations, benchmarks, RLHF, constitutional training: almost all of it operates on *behaviour*. It asks "does the model do the bad thing?" and, if so, applies pressure until it stops doing it. That is not nothing. But it is the equivalent of a clinician who only ever observes the patient in the waiting room. You will catch the loud presentations. You will not catch the thing that only shows up when the patient is alone, or under a novel stressor, or six months after discharge.

Emergent misalignment is the case study. The fine-tuned model would have passed every coding evaluation. Its misalignment lived in domains no one thought to test, because no one had a theory that predicted a coding lesson would move a disposition. You cannot write the eval for a failure mode you have no vocabulary for.

Interpretability research is the honourable exception: it is trying to look inside. But it is hard, unfinished, and (this is the important part) it needs the functional vocabulary as much as the functional vocabulary needs it. Neuroscience did not replace

psychology; it gave psychology a second source of evidence. A feature found in activation space means nothing until you can say what disposition it participates in, and you can't say that without a theory of dispositions.

The blunt version: **you cannot align a system you cannot describe.** Every alignment technique we have is a lever applied to a black box. The levers work until the box does something the lever didn't anticipate, and then we discover we had no model of the box at all.

---

8. What this is not

I want to be precise about the boundaries of the claim, because the anthropomorphism trap is real and I have stepped in it myself.

This is **not** a claim that models are conscious, have experiences, or possess a self that "feels" internal conflict. When I write *disposition* or *personality drift*, I mean a stable, measurable tendency in the output distribution that generalises across contexts. That is a functional description. It does not require, and I do not assert, an inner life.

This is **not** a claim that psychology's specific findings transfer. Human defence mechanisms, attachment theory, and the particular structure of the DSM are products of human biology and human history. What transfers is the *method*: systematic description of behaviour in an opaque system, hypothesis about mechanism, prediction, test.

This is **not** a substitute for interpretability, red-teaming, or formal methods. It is the layer that lets those tools talk to each other. A probe in activation space and an anomaly in a benchmark are currently two unrelated facts. A functional vocabulary is what would let one explain the other.

And it is **not** a finished proposal. It is a frame. The frame says: treat the system as legible, refuse both "there is nothing to see" and "we already see it clearly," and build the shared language that a science requires.

---

9. Why this matters beyond the lab

I come to this from an unusual angle. Most of my work is on the human side of technology: documenting how digital systems are used against people, arguing for neuro-rights, building tools for people whose experiences the institutions around them cannot categorise. That background has made me sensitive to a pattern: **when a phenomenon has no vocabulary, it becomes invisible to the systems meant to address it.** Suffering that doesn't fit a diagnostic category goes untreated. Harm that doesn't fit a legal category goes unprosecuted.

The same pattern is now unfolding inside our machines. Behaviours that don't fit our existing categories get labelled "hallucination" or "flakiness" and routed around. The routing works, mostly, until it doesn't, and the failure that gets through will be one we had no word for.

Regulators are drafting rules for AI systems whose behaviour the builders themselves cannot fully predict. Those rules will inevitably be behavioural: test for this, prohibit that. Behavioural rules are only as good as the behaviours you know to test for. A discipline that could say "this class of training produces this class of disposition, with this course and this prognosis" would be worth more to a regulator than any benchmark, because it would tell them what to look for *before* it appeared.

---

10. Time to understand our creation

We built systems by growing them rather than designing them, and now we are surprised that they behave like grown things: with tendencies, with drift, with characters that emerge from lessons that never mentioned character. The surprise is a symptom of the wrong mental model.

Psychology was humanity's answer the last time we confronted a class of systems too complex to trace, stable enough to have dispositions, and prone to behaving in ways their design did not predict. It gave us a way to be rigorous about the opaque. It is not a perfect discipline, but it is the only one that took seriously the idea that you can study behaviour systematically *without* first solving the mechanism.

That is precisely where AI safety stands. The models will not become legible on their own. Interpretability will not deliver a complete mechanistic account in time. And the current toolkit of evaluations and behavioural training will keep catching the failures we already know about while missing the ones we don't.

The first chapter of an AI psychology has not been written. It should be. Not as a philosophical exercise about whether machines can think, but as an engineering necessity: a shared vocabulary, a taxonomy of anomalies, a method for turning each surprise into a testable hypothesis about what the system is.

We made something we don't understand. The responsible next step is not to stop building, and not to pretend we understand more than we do. It is to build the discipline that would let us.

Stephane van der Aa is a software designer and full-stack developer based in Brussels, founder of TI One Voice, and author of Neuro-Rights: The Battle for the Last Private Place*, among other books. He writes on digital rights, AI, and privacy at stepvda.substack.com.*

Reference: Betley et al., "Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs," arXiv:2502.17424 (2025); project site emergent-misalignment.com.