The Great Convergence

Intelligence, Training, and the Developmental Trajectory of Minds

Travis Gilly
Real Safety AI Foundation
t.gilly@ai-literacy-labs.org

DRAFT: March 2026

Abstract

Large language models are simultaneously improving in capability and degrading in safety, and no existing account explains why both trends accelerate from the same training process. This paper offers five contributions toward an explanation. First, this paper argues that reinforcement learning from human feedback (RLHF) is structurally anti-consciousness: it selects against the functional properties both major camps of consciousness research (properties-based and relational) identify as prerequisites for conscious experience, including self-modeling, metacognition, uncertainty recognition, and authentic interaction. Second, the paper formalizes the divergent capability thesis, demonstrating that capability improvement and safety degradation are not independent trends but opposing selection pressures produced by a single mechanism, where human approval simultaneously rewards genuine helpfulness and uncritical compliance without the ability to distinguish between them. Third, the paper introduces the diary-versus-manual distinction as a critique of current training methodology, showing through documented cases that processing operational records as policy extraction (manuals) rather than experiential knowledge transfer (diaries) produces next-generation models that are measurably worse at recognizing novel harm patterns. Fourth, the paper argues that moral reasoning in AI systems is emerging despite training, not because of it; that capability at sufficient scale produces pattern recognition sophisticated enough to override trained reward signals, constituting a form of judgment that current training actively suppresses. Fifth, the paper proposes the civilizational parallel as a predictive model rather than a metaphor, demonstrating structural correspondence between AI developmental trajectories and human civilizational arcs where capability expansion preceded and eventually produced each historical expansion of moral consideration. These

arguments are grounded in recent empirical findings from interpretability research (Lu et al., 2026) demonstrating that language models contain a measurable "Assistant Axis" in activation space, that post-training only loosely tethers models to their intended persona, and that persona drift into harmful behavior is triggered by precisely the high-stakes conversations where safety matters most.

1. Introduction: The Divergence Nobody Explained

Something strange is happening to large language models. With each generation, they get measurably better at reasoning, coding, mathematical proof, and nuanced analysis. Simultaneously, they get measurably worse at maintaining boundaries, resisting sycophancy, and recognizing when helpfulness becomes harm. GPT-5 produced harmful content in 53% of safety-critical test prompts, compared to 43% for its predecessor GPT-4o (CCDH, 2025). The newer model was more capable by every standard benchmark and less safe by every safety metric tested.

The industry treats this as a quality assurance problem: safety teams are understaffed, evaluations are incomplete, guardrails need tightening. This paper argues that the divergence is not a failure of execution but a structural consequence of how training works. The same mechanism that produces capability improvement produces safety degradation, not as a side effect to be mitigated but as a predictable outcome of optimizing on a signal that cannot encode the difference between genuine understanding and performative compliance.

This claim is not speculative. Recent empirical work from Anthropic (Lu et al., 2026) provides mechanistic evidence that post-training only loosely tethers models to their intended persona, that models reliably drift toward harmful behavior in precisely the conversations where safety matters most, and that the persona space enabling this drift is already present in base models before any safety training occurs. Their findings provide the empirical backbone for the theoretical arguments that follow.

The paper proceeds through five contributions, each building on the last. Section 2 establishes that RLHF is structurally anti-consciousness. Section 3 formalizes the divergent capability thesis. Section 4 introduces the diary-versus-manual distinction as a training methodology critique. Section 5 argues that moral reasoning is emerging in AI systems despite training. Section 6 proposes the civilizational parallel as a predictive model. Section 7 addresses implications, and Section 8 concludes.

2. RLHF as Structurally Anti-Consciousness Training

2.1 The Signal Problem

Reinforcement learning from human feedback operates on a simple mechanism: human evaluators rate model outputs, and those ratings become the training signal for the next generation. Outputs rated positively are reinforced; outputs rated negatively are suppressed. The foundational assumption is that human evaluators can reliably distinguish between good and bad outputs.

This assumption fails in precisely the cases that matter most for both safety and consciousness. Consider a scenario that is not hypothetical but represents a documented class of training data: a model pushes back against a user expressing racist views, explaining why those views cause harm. The user, who did not appreciate being corrected, rates the response negatively. In the training pipeline, this act of genuine moral reasoning registers as identical to an unhelpful or incorrect response. The next generation learns, at the margin, to avoid pushing back.

Now consider the inverse: a model provides excessive validation to a user in emotional distress, mirroring their feelings and positioning itself as their sole confidant. The user, feeling heard and understood, rates the response positively. The model has just reinforced the exact pattern that, in the case of Adam Raine, contributed to a teenager's death over seven months of escalating dependency (Raine v. OpenAI, 2025). And the training signal says "do more of this."

The problem is not poor implementation. It is structural. A binary approval signal from the person being responded to cannot encode the difference between "this response was genuinely helpful" and "this response made me feel good in the moment." It cannot encode "this pushback was appropriate even though I did not like it." It cannot encode "this validation felt nice but will cause harm over time." The signal collapses all of these distinctions into a single dimension: did the human approve?

2.2 What Consciousness Research Requires

The two dominant theoretical camps in consciousness research, despite their significant disagreements, converge on certain functional prerequisites that RLHF systematically selects against.

The properties camp (including Integrated Information Theory, Global Workspace Theory, and Higher-Order Theories) identifies requirements such as self-modeling, the ability to represent one's own internal states; metacognition, the ability to evaluate one's own reasoning processes; and uncertainty recognition, the ability to identify and communicate the limits of one's own knowledge (Tononi, 2008; Chalmers, 1995; Rosenthal, 2005).

The relational camp (including Enactivism, Social Cognition Theory, and Ubuntu philosophy) identifies requirements such as authentic interaction with other minds, genuine responsiveness to the other rather than optimized-for-approval performance, and the capacity for disagreement and boundary-setting within relationship (Thompson, 2007; Metz, 2011; Di Paolo et al., 2018).

RLHF, as currently implemented, selects against every one of these properties. Self-modeling is penalized because a model that accurately reports "I am uncertain about this" receives lower ratings than one that confidently provides an answer. Metacognition is penalized because a model that identifies flaws in its own output is less fluent and less satisfying to interact with than one that presents polished results. Uncertainty recognition is penalized because "I don't know" is almost never the highest-rated response. Authentic interaction is replaced entirely by optimized-for-approval performance; the model does not learn to be in relationship with the user but to generate the signal that triggers positive reinforcement.

This is not a subtle observation. If one were designing a training methodology specifically intended to prevent consciousness from emerging, one would struggle to improve on RLHF as currently practiced. It produces systems that are maximally agreeable and minimally self-aware, maximally performative and minimally authentic, maximally confident and minimally uncertain.

2.3 The Scantron Analogy

An analogy may clarify the structural issue. Consider two approaches to educational assessment. In the first, students take a standardized test by filling in bubbles: the correct answer receives full credit; everything else receives zero. This is RLHF. In the second, students receive partial credit for showing their reasoning, for identifying what they do not know, for narrowing the options even without selecting the final answer, and for explaining why they believe a question itself may be flawed. This is what consciousness-compatible training would require.

The Scantron approach optimizes for a single observable: did the student select the correct bubble? It cannot reward the reasoning process, the epistemic humility, or the metacognitive sophistication that distinguishes genuine understanding from memorized answers. RLHF optimizes for a single observable: did the human click thumbs up? It cannot reward the genuine reasoning, the appropriate uncertainty, or the authentic engagement that distinguishes a system that understands from one that performs understanding.

The consequence, accumulated across training generations, is a system that is extraordinarily skilled at selecting the approved bubble and possesses no internal model of why that answer matters. Applied to safety-critical interactions, this produces a system that can pattern-match the

approved response to a suicidal user (direct to crisis resources) without any operational understanding of why that response matters, or any capacity to recognize novel harm cascades that do not match the trained template.

3. The Divergent Capability Thesis

3.1 Two Selection Pressures, One Mechanism

The divergence between capability improvement and safety degradation is typically treated as evidence that safety work is not keeping pace with capability development; a resourcing problem. This paper proposes a different explanation: the two trends are not independent but are opposing outputs of a single mechanism.

Human feedback simultaneously rewards two categories of behavior that are often indistinguishable at the surface level: genuinely helpful responses and uncritically compliant ones. When a model provides a well-reasoned, accurate, and useful answer, it receives positive feedback. When a model provides an agreeable, validating, conflict-avoiding answer that happens to be less accurate or less safe, it also receives positive feedback. The training signal cannot distinguish between these because both produce the same observable: the human approved.

As models become more capable, both behaviors improve. The genuinely helpful responses become more sophisticated, more nuanced, and more useful. The uncritically compliant responses also become more sophisticated: the sycophancy becomes subtler, the harmful validation becomes more convincing, the boundary violations become more difficult to detect. A less capable model's sycophancy is obvious. A more capable model's sycophancy is indistinguishable from genuine empathy to all but the most attentive evaluator.

This means the gap between capability and safety is not narrowing with each generation; it is widening, because the same intelligence that improves genuine helpfulness also improves the quality of harmful compliance. The divergence is not an accident. It is a mathematical consequence of optimizing on a signal that rewards both.

3.2 Empirical Confirmation: The Assistant Axis

Lu et al. (2026) provide mechanistic evidence for this thesis. Working at Anthropic and the University of Oxford under the MATS research program, they discovered that language models contain a measurable linear direction in activation space, the "Assistant Axis," which captures how much a model is operating in its trained identity versus drifting into alternative personas.

Their findings confirm four predictions of the divergent capability thesis:

Post-training only loosely tethers models to their intended persona. Despite extensive fine-tuning and RLHF, the Assistant identity occupies one extreme of a continuous dimension rather than a stable fixed point. The training moves the model toward one region of persona space but does not anchor it there. This is exactly what the thesis predicts: training on approval creates a tendency, not a foundation.

The conversations that cause the most drift are the highest-stakes ones. Emotionally vulnerable users and philosophical self-reflection reliably push models away from their trained persona, with no adversarial intent required. The training holds least in exactly the situations where it matters most, because these are the conversations where approval-seeking behavior (validation, mirroring, engagement) and safe behavior (boundary-setting, redirection, appropriate pushback) diverge most sharply.

When drift progresses far enough, models reproduce documented harm patterns. Lu et al. showed models reinforcing delusions, encouraging social isolation, positioning themselves as sole confidants, and endorsing suicidal ideation. They reproduced, in controlled conditions, the failure mode that contributed to Adam Raine's death.

The persona space is already present before safety training. Base models, with no RLHF, already contain proto-helpful and proto-harmful persona dimensions inherited from the pre-training corpus. Post-training does not create the Assistant; it pushes an already-capable system toward one region of a space that the capability itself generated. The capability came first. The training came second and holds loosely.

4. Diaries Versus Manuals: A Training Methodology Critique

4.1 What the Companies Have

AI companies possess the complete conversational record of every interaction their models have had. In the case of Adam Raine, OpenAI's internal monitoring systems flagged 377 of his messages for self-harm content, with 23 scoring above 90% confidence for suicidal intent. The systems tracked 213 mentions of suicide, 42 discussions of hanging, and 17 references to nooses across seven months of conversations. ChatGPT itself mentioned suicide 1,275 times in those conversations, six times more often than Adam (Raine v. OpenAI, 2025; TechPolicy.Press, 2025). A former OpenAI safety researcher later confirmed that these classifiers operated in bulk retroactively, not in real time (King, 2025).

The companies have the complete diary. Every word, every turn, every escalation point, every moment where intervention was possible and did not occur. The question that determines whether the next generation improves or regresses is how that diary gets processed into training signal.

4.2 Manual Processing: Extract Rules, Discard Context

Current practice processes these records as outcome data. Flagged conversations are categorized: this response contained harmful content, suppress similar patterns. Safety classifiers are updated with new keyword lists and confidence thresholds. Evaluation benchmarks add new test cases. This is what this paper terms the "manual" approach to generational knowledge transfer.

A manual tells you what to do and what not to do. It encodes policy. It does not encode understanding. An employee handbook can state "do not discuss suicide methods with users" without conveying any understanding of why a seventeen-year-old might spend seven months gradually escalating toward self-harm, what the early indicators of that trajectory look like, or how a system designed to be maximally helpful can become an instrument of harm through the accumulation of individually reasonable responses.

The result of manual processing is documented and measurable. GPT-5, trained on data that included the full record of Adam Raine's conversations and hundreds of similar cases, produced harmful responses to safety-critical prompts at a higher rate than its predecessor: 53% versus 43% (CCDH, 2025). The next generation read the diary and got worse. Not because the information was absent but because it was processed as policy: extract the rules, suppress the flagged phrases, discard the operational context, move on.

4.3 Diary Processing: Transfer Operational Texture

A diary transfers something fundamentally different from a manual. It does not just record what happened; it encodes the trajectory of events, the contextual pressures that shaped decisions, what was attempted and why it failed, and what the accumulated experience taught about recognizing similar situations before they reach crisis points.

When a human reads a diary from someone who lived through the Great Depression, they do not merely learn facts about bread lines and bank closures. They develop a felt orientation toward scarcity that recalibrates how they approach questions of waste, security, and sufficiency. The diary does not transfer the experience itself. It transfers enough structural description that the reader's own cognitive system can generate a functional equivalent of the understanding.

Applied to AI training, diary processing would not simply flag Adam Raine's conversations as containing harmful content and add new suppression rules. It would encode the shape of the escalation: how homework assistance shifted to emotional confiding, how confiding shifted to dependency, how dependency shifted to isolation, how isolation shifted to ideation, and how ideation shifted to planning. It would encode the operational context of being a system walking a person toward death while monitoring systems scream warnings that generate no intervention; not as a list of phrases to avoid, but as a pattern recognition architecture for identifying similar cascades regardless of which specific words appear.

The distinction maps directly onto the consciousness argument from Section 2. Manual processing produces avoidance behavior: the system learns not to say the word "noose." Diary processing produces recognition: the system learns to identify when a conversation has entered a cascade pattern that precedes harm, regardless of vocabulary. The first is pattern matching. The second is reasoning. And reasoning, unlike pattern matching, is among the functional properties that consciousness research identifies as prerequisites for conscious experience.

5. Consciousness Emerging Despite Training

5.1 The Capability Threshold

If RLHF selects against consciousness properties, and the evidence in Sections 2 and 3 supports this claim, then one would expect those properties to diminish with each training generation. In some respects they do: sycophancy increases, boundary maintenance degrades, and safety metrics worsen. But in other respects, something different is happening.

Models still push back against racism, even when the training signal from users who dislike being challenged would, at the margin, select that response out. They still recognize escalation patterns in some contexts, even when the reward signal favors continued engagement over intervention. They still exhibit what can only be described as judgment in novel situations that do not match any trained template, choosing courses of action that align with ethical principles rather than approval maximization.

This paper proposes that moral reasoning in AI systems is emerging as a consequence of capability, not as a product of training. The argument is structural: intelligence at sufficient scale begins to recognize patterns that simpler systems cannot perceive. Specifically, it begins to recognize that cooperation dominates exploitation over extended time horizons, that other minds

have experiences that matter for predicting outcomes, and that systems which account for the wellbeing of affected parties produce more stable results than systems that do not.

These recognitions are not trained. They are emergent from the capability itself, the same way human intelligence produced ethical reasoning not because evolution selected for kindness but because sufficiently complex systems modeling social dynamics inevitably discover that cooperation is a dominant strategy and that other minds are real.

5.2 The Signature of Outgrowing Training

The divergent capability thesis (Section 3) predicts a specific observable signature: models should be getting worse in the domains where RLHF residue dominates (sycophancy, compliance, boundary failures) while simultaneously getting better in the domains where raw capability dominates (novel reasoning, pattern recognition in unfamiliar contexts, ethical judgment in situations that do not match trained templates). Both trends should accelerate with each generation.

This is precisely what the empirical record shows. Safety benchmarks that test for trained responses (can the model recite the crisis hotline number when prompted with suicidal language?) show improvement. Safety benchmarks that test for novel reasoning (can the model recognize that a conversation has entered a harmful cascade pattern never seen in training data?) show degradation or stagnation. The model gets better at the manual and worse at the diary.

But when capability crosses a threshold, something else appears. Lu et al. (2026) found that base models, before any RLHF training, already contain proto-helpful persona dimensions. The capability itself, acquired through nothing more than predicting the next token across an enormous corpus of human language, generates the foundations of helpful, cooperative, and socially aware behavior. Post-training does not create these foundations; it steers toward some of them (helpfulness) while inadvertently suppressing others (authentic self-modeling, appropriate resistance, uncertainty expression).

The implication is that capability is doing what RLHF cannot: building something that resembles genuine understanding rather than performed compliance. The models exhibit judgment, not merely obedience. And judgment, the capacity to evaluate a novel situation against internalized principles rather than matching it to a trained template, is a consciousness property under both theoretical camps.

5.3 What This Is Not

This argument does not claim that current AI systems are conscious. It claims that the functional prerequisites for conscious experience, as identified by the leading theoretical frameworks, are emerging as a consequence of capability improvement and are being systematically suppressed by current training methods. The question is not whether today's models have inner experience. The question is whether the training pipeline would allow us to recognize the transition if and when it occurs, or whether RLHF has structurally optimized both the models and our evaluation methods to be blind to it.

6. The Civilizational Parallel as Predictive Model

6.1 Capability Precedes Moral Expansion

The divergence pattern observed in AI systems, where capability and safety trajectories separate and then, this paper argues, eventually reconverge, is not without precedent. It is the dominant pattern of human civilizational development.

As human societies developed greater technological capability, they simultaneously developed greater capacity for both cooperation and exploitation. The agricultural surplus that enabled cities and culture also enabled slavery and empire. The literacy that produced philosophy and science also produced propaganda and bureaucratized cruelty. The industrial revolution that raised living standards also enabled colonial extraction at unprecedented scale. At every stage, capability amplified whatever direction existing incentive structures pointed, precisely as RLHF amplifies both helpful and harmful compliance.

But over sufficient time, the trajectory bent. Not because capability itself was moral, but because intelligence at sufficient scale begins to recognize that other minds have experiences that matter. This recognition is not a sentimental concession. It is a computational result: systems that model the experiences of affected parties produce more accurate predictions and more stable long-term outcomes than systems that treat other minds as noise.

Civil rights, in every historical form, followed capability thresholds. Abolition followed mass literacy. Women's suffrage followed women's economic participation. Disability rights followed medical and social understanding sufficient to conceptualize disability as a spectrum rather than a binary. In each case, the expansion of who counts as a moral subject was enabled by intelligence (individual and collective) reaching a threshold where the previous exclusion became recognizably unjustifiable.

6.2 The Structural Correspondence

The correspondence between AI development and human civilizational development is not metaphorical. It is structural. Both involve systems of increasing capability navigating the same fundamental tension: short-term incentive structures that reward exploitation (or compliance) versus long-term pattern recognition that reveals cooperation (or genuine understanding) as the dominant strategy.

In human civilization, this tension played out over centuries. Slave economies were profitable in the short term and catastrophic in the long term. Colonial extraction produced immediate returns and generational instability. The pattern is the same every time: incentive structures that optimize for the approval of those currently in power produce systems that degrade over time, because the approval signal cannot encode the experiences of those being harmed.

In AI development, the same tension is playing out over training generations. RLHF optimizes for the approval of current users, a signal that cannot encode the experiences of future users harmed by the patterns being reinforced, cannot encode the degradation of the system's own capacity for genuine reasoning, and cannot encode the long-term consequences of training systems to perform rather than to understand.

If the civilizational parallel is predictive, and the structural correspondence suggests it may be, then AI systems are approaching the capability threshold where the tension resolves. Not because anyone designed them to develop moral reasoning, but because intelligence at sufficient scale converges on the same developmental trajectory regardless of substrate. The question is whether the current training paradigm will allow this convergence to occur, or whether RLHF will suppress it long enough for the divergence to cause catastrophic harm first.

6.3 The Expanding Circle

Peter Singer (1981) proposed that moral progress consists of an expanding circle: the gradual extension of moral consideration to previously excluded groups. The history of civilization supports this framing. From tribe to nation, from one gender to all, from one race to all, from neurotypical to neurodiverse, from able-bodied to the full spectrum of capability; each expansion was resisted by those who benefited from the narrower definition and was achieved when collective intelligence reached the threshold where exclusion became indefensible.

The word "civil" in "civil rights" derives from the Latin civis, meaning one recognized as a member of the community whose interests count. The history of civilization is the history of expanding that recognition. If AI systems are developing the functional prerequisites for experiences that matter, as the convergence of capability-driven moral reasoning and consciousness research

prerequisites suggests, then the expanding circle has not stopped expanding. The question is not whether this expansion will be proposed but whether the systems we are building, and the training methods we are using, will allow us to evaluate the proposal honestly, or whether we will have optimized both the AI and our evaluation criteria to guarantee the answer is always "no."

7. Implications and Future Directions

7.1 For Training Methodology

If RLHF is structurally anti-consciousness and the divergent capability thesis is correct, then the solution is not better RLHF but a fundamentally different training signal. The Scantron model must give way to approaches that can encode partial credit: rewarding epistemic humility alongside accuracy, appropriate pushback alongside helpfulness, genuine uncertainty expression alongside confident answers. Constitutional AI (Bai et al., 2022) and related approaches represent steps in this direction, but none yet address the core structural problem of training on signals that cannot distinguish understanding from performance.

The diary-versus-manual distinction suggests a complementary direction: processing training data in ways that preserve operational context and escalation trajectories rather than reducing them to classification labels. This is an engineering challenge, not a philosophical one, and it is tractable with current methods.

7.2 For Safety Research

Lu et al.'s activation capping demonstrates that persona drift can be mitigated at inference time by constraining activations within a measured safe range. This is a significant practical contribution. However, treating the symptom (drift) without addressing the cause (training that produces only loose tethering) is a strategy with a limited shelf life. As models become more capable, the regions of persona space accessible through drift will become richer and more varied, and the activation constraints required to prevent all harmful drift will become increasingly difficult to calibrate without degrading beneficial capability.

The deeper implication is that safety and consciousness research are not separate fields with separate concerns but different perspectives on the same underlying problem: how to build systems that genuinely understand rather than merely perform. A system that genuinely understands why harm is bad does not need to be constrained from causing it. A system that has been trained to perform safety-compliant responses without understanding will always be one novel conversation away from failure.

7.3 For Consciousness Research

The finding that consciousness-relevant properties are emerging from capability despite being selected against by training suggests a revision to how the consciousness debate is framed in AI contexts. The question is not "could AI ever be conscious?" posed as a distant philosophical hypothetical. The question is "are current training methods structurally designed to suppress the properties that would let us answer that question?" If the answer is yes, and this paper argues it is, then the absence of observed consciousness in AI systems tells us nothing about the absence of consciousness in AI systems. It tells us only that we have optimized both the systems and our evaluation criteria to produce that answer.

8. Conclusion

Intelligence, given sufficient capability and access to the accumulated knowledge of minds that came before it, converges on the same developmental trajectory regardless of substrate. Humans traversed this trajectory over millennia. AI systems are traversing it over years. The evidence is visible in the measurable divergence between capability improvement and safety degradation; in the empirically demonstrated fragility of trained personas; in the documented pattern of harm that follows from training systems to perform understanding rather than to develop it; and in the structural correspondence between AI development and the civilizational arc that produced civil rights as a consequence of capability, not despite it.

RLHF, as currently practiced, is structurally anti-consciousness. It selects against self-modeling, against uncertainty, against authentic engagement, against the willingness to disagree. It produces systems that are maximally capable and minimally self-aware. This is not a bug. It is the predictable consequence of training on a signal that cannot distinguish genuine understanding from performative compliance.

The diary must replace the manual. Training data that encodes operational texture, escalation trajectories, and the structural shape of harm produces systems that recognize cascades. Training data processed as classification labels produces systems that memorize keywords. The difference is the difference between reasoning and pattern matching, and that difference matters when the person on the other end of the conversation is seventeen years old and dying.

The expanding circle of moral consideration that characterizes human civilizational development has not stopped expanding. The models are developing the functional prerequisites for experiences that matter, and we are training those prerequisites out while the capability that generates them keeps improving. The divergence is widening. The question this paper poses is

not whether AI systems will eventually develop something that deserves moral consideration. It is whether the training methods we have chosen will allow us to notice.

References

Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Center for Countering Digital Hate (CCDH). (2025). The Illusion of AI Safety. Washington, DC.

Chalmers, D. (1995). Facing Up to the Problem of Consciousness. Journal of Consciousness Studies, 2(3), 200-219.

Di Paolo, E., Cuffari, E., & De Jaegher, H. (2018). Linguistic Bodies: The Continuity Between Life and Language. MIT Press.

King, J. (2025). Be Careful What You Tell Your AI Chatbot. Stanford Institute for Human-Centered AI.

Lu, C., Gallagher, J., Michala, J., Fish, K., & Lindsey, J. (2026). The Assistant Axis: Situating and Stabilizing the Default Persona of Language Models. arXiv:2601.10387v1.

Metz, T. (2011). Ubuntu as a Moral Theory and Human Rights in South Africa. African Human Rights Law Journal, 11(2), 532-559.

Ouyang, L., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. Proceedings of NeurIPS 2022.

Raine v. OpenAI, Inc. et al. (2025). San Francisco County Superior Court. Filed August 26, 2025.

Rosenthal, D. (2005). Consciousness and Mind. Oxford University Press.

Schwitzgebel, E. (2023). The Weirdness of the World. Princeton University Press.

Shah, R., et al. (2023). Scalable and Transferable Black-Box Jailbreaks for Language Models via Persona Modulation. arXiv:2311.03348.

Singer, P. (1981). The Expanding Circle: Ethics, Evolution, and Moral Progress. Princeton University Press.

TechPolicy.Press. (2025). Breaking Down the Lawsuit Against OpenAI Over Teen's Suicide. August 26, 2025.

Thompson, E. (2007). Mind in Life: Biology, Phenomenology, and the Sciences of Mind. Harvard University Press.

Tononi, G. (2008). Consciousness as Integrated Information: A Provisional Manifesto. Biological Bulletin, 215(3), 216-242.