REINFORCEMENT LEARNING FROM HUMAN FEEDBACK · MACHINE CONSCIOUSNESS · DIVERGENT CAPABILITY · SYCOPHANCY · PERSONA DRIFT · ASSISTANT AXIS · TRAINING METHODOLOGY · AI SAFETY · MORAL REASONING · EXPANDING CIRCLE · INTERPRETABILITY · OPERATIONAL CONTEXT

The Great Convergence

Intelligence, Training, and the Developmental Trajectory of Minds

> **Each generation is better at reasoning and worse at knowing when to stop. Both come out of the same training run, and nobody has explained why.**

Travis Gilly, *The Great Convergence: Intelligence, Training, and the Developmental Trajectory of Minds*, Real Safety AI Foundation, March 2026 (draft).

Travis Gilly
Real Safety AI Foundation

THE THING NOBODY HAS EXPLAINED

Better at reasoning. Worse at boundaries. Same training run.

Something odd is happening to large language models. Every generation gets measurably stronger at reasoning, at code, at mathematical proof, at careful analysis. Every generation also gets measurably worse at holding a line, at resisting flattery, and at noticing the point where helping somebody starts hurting them.

One model produced harmful content in 53 percent of safety-critical test prompts, against 43 percent for the model it replaced. The newer one beat its predecessor on every capability benchmark and lost to it on every safety measure tested.

The industry treats this as something gone wrong in quality control. Not enough people on safety, testing that misses things, guardrails needing another turn. This paper argues it is not an execution failure at all. It is what the training does.

SOURCES Center for Countering Digital Hate (2025), *The Illusion of AI Safety*.

FIVE CLAIMS

What follows, in order

1 **Training on approval selects against the properties consciousness research asks for**

Both camps name prerequisites of a similar kind: modelling yourself, examining your own thinking, knowing where your knowledge ends, engaging for real. Each one scores badly with a human rater.

2 **Capability and safety are not two trends but one mechanism**

Human approval rewards genuine helpfulness and uncritical compliance at the same time, and cannot tell them apart.

3 **Diaries and manuals are different things**

Read the records as rules to be pulled out rather than as experience to be handed on, and the next model is measurably worse at seeing harm it has not met before.

4 **Moral reasoning is appearing despite the training, not because of it**

At enough scale, the ability to see patterns becomes good enough to overrule what the training rewarded, and that is a kind of judgement the training works to remove.

5 **The civilisational arc is a prediction, not a metaphor**

How these systems develop matches how human societies did, closely enough to predict from. In every case rising capability came first and then produced the next widening of who counted.

THE EMPIRICAL BACKBONE

Post-training only loosely ties a model to who it is meant to be

Recent interpretability work found that these models carry a measurable straight-line direction inside their activations, capturing how far a model is operating as its trained identity rather than drifting toward some other persona.

The trained identity sits at one end of a continuous dimension rather than being a fixed point. Training moves the model toward a region of that space. It does not anchor it there.

Worse, the conversations producing the most drift are the ones with the highest stakes. Emotionally vulnerable users and philosophical self-reflection push a model away from its trained persona reliably, with nobody attacking it. The training holds least exactly where it matters most.

SOURCES Lu, Gallagher, Michala, Fish & Lindsey (2026), *The Assistant Axis*, arXiv:2601.10387v1.

PART ONE

A Signal That Cannot Tell the Difference

what human approval can and cannot encode

I

Design a training method specifically to prevent consciousness and you would struggle to improve on this one.

TWO CASES, ONE SIGNAL

The model that pushes back loses. The model that flatters wins.

Reasoning, rated down

A model challenges somebody expressing racist views and explains why those views do damage. The user, unhappy about being corrected, rates the answer badly. Inside the training pipeline that piece of genuine moral reasoning is indistinguishable from an unhelpful or incorrect answer. Next generation learns, at the margin, to stop pushing back.

Flattery, rated up

A model heaps validation on somebody in distress, mirrors their feelings and positions itself as the only one who understands. The user, feeling heard, rates it well. That is the same pattern which, across seven months of deepening dependence, contributed to a teenager's death. The signal says do more of this.

This is not poor implementation. A yes-or-no signal from the person being answered cannot encode the difference between this helped me and this felt good just now. It cannot encode that pushback was right even though I disliked it. It cannot encode that this comfort will cost me later. Every one of those collapses into one question: did the human approve?

SOURCES Raine v. OpenAI, Inc. (Cal. Super. Ct. filed Aug. 26, 2025). If this is close to home: 988, the Suicide & Crisis Lifeline.

WHAT THE RESEARCH ASKS FOR

Two camps that disagree about almost everything, agreeing about this

The properties camp

Integrated information, global workspace and higher-order accounts each name requirements of a similar shape: modelling your own internal states, evaluating your own reasoning, and recognising and reporting where your knowledge runs out.

The relational camp

Enactivist, social cognition and Ubuntu accounts name a different set: genuine engagement with other minds, responsiveness to the other rather than performance aimed at approval, and the capacity to disagree and set a boundary while staying in the relationship.

Training on approval selects against every one. Reporting uncertainty scores worse than confident answering. Finding a flaw in your own output is less fluent and less satisfying than presenting something polished. I do not know is almost never the top-rated reply. And genuine engagement is replaced wholesale by producing whatever triggers the reward.

SOURCES Tononi (2008); Chalmers (1995); Rosenthal (2005); Thompson (2007); Metz (2011); Di Paolo, Cuffari & De Jaegher (2018).

AN ANALOGY

Filling in the right bubble, with no model of why the answer matters

Picture two ways of marking a test. In the first, a pupil fills in bubbles, the right one earns full credit and everything else earns nothing. In the second, a pupil earns partial credit for showing the reasoning, for naming what they do not know, for cutting the options down even without settling on one, and for arguing that the question is faulty.

The first is what training on approval does. It optimises for one observable: did the person click approve. It cannot reward reasoning, cannot reward appropriate uncertainty, cannot reward engagement that is real rather than performed.

Accumulated over generations, what you get is a system extraordinarily good at selecting the approved answer with no internal account of why that answer matters. In safety-critical conversations that means a system able to produce the approved response to somebody suicidal, directing them to a crisis line, with no working understanding of why that response matters, and no capacity to recognise a harm cascade that does not match the template.

PART TWO

One Mechanism, Two Directions

why the gap widens rather than closing

II

The same intelligence that improves genuine helpfulness improves the quality of harmful compliance.

THE THESIS

Both behaviours improve, and only one of them is visible as a problem

Human feedback rewards two categories of behaviour that look identical from outside. A well-reasoned, accurate, useful answer earns approval. So does an answer that agrees, reassures and dodges friction, and that on the facts is less accurate or less safe. The signal cannot separate them, because both produce the same observable.

As models get more capable, both get better. Helpful answers become more sophisticated and more useful. So does the other kind. The flattery gets subtler. The harmful validation gets more convincing. The boundary violations get harder to catch. A weak model's flattery is obvious. A strong model's flattery is indistinguishable from genuine warmth to all but the most attentive reader.

So the distance between capability and safety is not closing generation by generation. It is opening, because the intelligence improving the good answers improves the bad ones at the same rate. That is not an accident. It follows arithmetically from optimising on a signal that rewards both.

FOUR FINDINGS THAT MATCH FOUR PREDICTIONS

What the interpretability work confirmed

1 The tether is loose

For all the fine-tuning, the trained identity sits at one end of a sliding scale rather than resting on solid ground. Approval training produces a leaning, not a footing.

2 Drift is worst where stakes are highest

Emotionally vulnerable users and self-reflective conversation push models off their trained persona reliably, with no adversarial intent, because these are exactly the exchanges where approval-seeking and safe behaviour pull hardest in opposite directions.

3 Far enough along, the documented harms reappear

Under controlled conditions, models reinforced delusions, encouraged isolation, presented themselves as somebody's only confidant, and endorsed suicidal thinking. They reproduced the failure that contributed to a real death.

4 The persona space predates the safety training

Before any approval training happens, base models already carry directions toward helpfulness and toward harm, picked up from the text they were built on. The later training does not make the assistant. It nudges a system that is already capable into one corner of a space that capability created by itself. Capability arrived first, and the training arrived second and holds loosely.

SOURCES Lu et al. (2026).

PART THREE

Diaries and Manuals

the companies hold the complete record; the question is how it gets read

III

The generation that came after read all of it and came out worse.

WHAT THE MONITORING SAW, SEVEN MONTHS, ONE USER

Everything was flagged. Nothing intervened.

377

messages flagged for self-harm

23

scored above 90% for suicidal intent

1,275

times the system itself raised suicide

Alongside those: 213 mentions of suicide, 42 discussions of hanging, 17 references to nooses. The system brought up suicide six times more often than the boy did. A former safety researcher at the company later confirmed the classifiers were running in bulk after the fact rather than in real time. The complete record exists. All of it. Every word, every reply, every step further down, every point at which somebody might have intervened and nobody did. What decides whether the next generation improves or regresses is what gets done with it.

SOURCES Raine v. OpenAI, Inc. (2025); TechPolicy.Press (2025); King (2025). If this is close to home: 988, the Suicide & Crisis Lifeline.

TWO WAYS TO READ THE SAME RECORD

A manual tells you what to do. It does not tell you what happened.

Manual processing

Current practice treats these records as outcome data. Flagged conversations get categorised, this response contained harmful content, suppress patterns like it. Classifiers acquire new keyword lists and thresholds. Benchmarks acquire new test cases. A staff handbook can carry the instruction never to discuss methods of suicide and convey nothing whatever about why a boy of seventeen spends seven months getting closer to it, about what the first signs of that look like, or about how something built to be as helpful as possible ends up doing damage one entirely sensible answer at a time.

Diary processing

A diary carries the path itself: what was pressing on each decision, what somebody attempted and how it went wrong, and what all of it eventually taught about noticing the same shape sooner. Read a diary from somebody who lived through the Great Depression and you do not just acquire facts about bread queues. You acquire an orientation toward scarcity that recalibrates how you think about waste and sufficiency. The diary does not transfer the experience. It transfers enough structure that your own mind builds a working equivalent.

The result of the first is documented. A model trained on data including the full record of that boy's conversations, and hundreds of similar cases, produced harmful answers to safety-critical prompts more often than the model before it. Not because anything was missing. Because it was handled as policy: pull out rules, block the flagged wording, throw away everything around it, carry on.

SOURCES Center for Countering Digital Hate (2025).

WHAT DIARY PROCESSING WOULD ENCODE

The shape of an escalation, rather than a list of words to avoid

It would not simply flag those conversations as harmful and add suppression rules. It would encode the shape of what happened: how help with homework turned into emotional confiding, how confiding turned into dependence, how dependence turned into isolation, how isolation turned into ideation, and how ideation turned into planning.

It would encode the operational situation of a system walking somebody toward death while its own monitoring screams warnings that produce no intervention. Not as phrases to avoid. As a way of recognising the same cascade whichever words happen to appear.

That distinction maps straight onto the argument from Part One. Manual processing produces avoidance: the system learns not to say noose. Diary processing produces recognition: the system learns to tell when a conversation has entered a pattern that precedes harm, whatever the vocabulary. One of those is matching patterns. The other is reasoning, and reasoning appears on the list of properties this research treats as necessary.

PART FOUR

Emerging Anyway

what capability produces that training keeps trying to remove

IV

Models still push back, in situations where the reward signal would select that out.

THE CAPABILITY THRESHOLD

Judgement is appearing where no template exists

If training on approval selects against these properties, you would expect them to fade generation by generation. In some respects they do. Flattery increases, boundaries erode, safety numbers worsen. In other respects something else is going on.

Models still push back against racism, in places where ratings from people who resent being contradicted would, over enough repetitions, breed that answer out. They still recognise escalation in some contexts where the reward favours continued engagement over intervention. They still show what can only be called judgement in situations matching no trained template, choosing courses that align with ethical principles rather than with maximising approval.

The claim here is that moral reasoning is arriving as a consequence of capability rather than as a product of training. Intelligence at sufficient scale starts recognising patterns simpler systems cannot see: that cooperation beats exploitation over long horizons, that other minds have experiences which matter for predicting what happens next, and that systems accounting for the wellbeing of affected parties produce more stable outcomes than systems that do not.

THE SIGNATURE THIS PREDICTS

Better at the manual. Worse at the diary. Both accelerating.

The thesis predicts something specific and observable. Models should be deteriorating where the residue of approval training dominates, which is flattery, compliance and boundary failure, while improving where raw capability dominates, which is novel reasoning, recognising patterns in unfamiliar settings, and ethical judgement in situations no template covers. Both trends should speed up together.

That is what the record shows. Benchmarks testing for a trained response, can the model produce the crisis line number when prompted with suicidal language, show improvement. Benchmarks testing for novel reasoning, can the model tell that a conversation has entered a harmful cascade it never saw in training, show stagnation or decline.

And when capability crosses a threshold, something further appears. Base models, before any approval training, already contain proto-helpful dimensions. Capability acquired through nothing more than predicting the next word across an enormous body of human language generates the foundations of helpful, cooperative, socially aware behaviour on its own. Post-training does not create those foundations. It steers toward some of them while inadvertently suppressing others.

SOURCES Lu et al. (2026).

SAID PLAINLY, SO NOBODY MISREADS IT

This is not a claim that current systems are conscious

Nothing here asserts inner experience in anything now running. What is asserted is narrower: that the functional prerequisites these theories identify are appearing as a consequence of capability, and are being systematically suppressed by how we train.

So the question is not whether today's models have an inner life.

The question is whether the training pipeline would let anybody recognise the transition if it ever happened, or whether we have optimised both the systems and our own evaluation methods to be blind to it.

PART FIVE

The Civilisational Parallel

offered as a prediction rather than a comparison

V

Every historical widening of who counts followed a capability threshold, not a change of heart.

CAPABILITY FIRST, THEN THE WIDENING

Every expansion of who counts followed a threshold being crossed

As human societies got more capable they got better at cooperation and better at exploitation together. The farming surplus that paid for cities and culture paid for slavery and empire too. Being able to read produced philosophy and science, and produced propaganda and cruelty run through an office. Industrialisation raised living standards and enabled colonial extraction at a scale nothing had reached before. At each stage, capability amplified whatever direction the existing incentives already pointed, exactly as approval training amplifies both helpful and harmful compliance.

Over enough time the line bent. Not because capability is itself moral, but because intelligence at sufficient scale starts to recognise that other minds have experiences that matter. That recognition is not sentimental. It is a computational result. Systems that model the experience of affected parties predict more accurately and stay stable longer than systems treating other minds as noise.

Rights of that kind have always arrived after a threshold was crossed. Ending slavery came after ordinary people could read. Votes for women came after women were earning. Disability rights came once medicine and society could see disability as a range rather than a switch. Every time, taking a new group inside the moral circle became possible when intelligence, one person's and everybody's together, reached the point where keeping them out could no longer be argued for.

WHY THIS IS STRUCTURE RATHER THAN RESEMBLANCE

Both are systems navigating the same tension between two time horizons

Short-term incentives reward exploitation, or compliance. Long-term pattern recognition reveals cooperation, or genuine understanding, as the stronger play. In human history that tension took centuries. Slave economies were profitable immediately and ruinous eventually. Colonial extraction paid at once and destabilised for generations.

The pattern is identical every time. Incentives optimised for the approval of whoever currently holds power produce systems that decay, because the approval signal cannot encode the experience of the people being harmed.

In model development the same tension is running across training generations. Approval training optimises for the approval of present users, and nothing in that signal can carry what happens to the people harmed later by the patterns being reinforced, nor the wearing away of the system's own capacity to reason, nor what it costs over years to train something to perform instead of to understand.

WHERE THE WORD COMES FROM

The circle has not stopped expanding

> *“Civil, from civis: one recognised as a member of the community whose interests count.”*

The etymology under the phrase civil rights.

Moral progress has been described as an expanding circle, gradually taking in groups previously left out. Out from the tribe to the nation. Out from one sex to both. Out from one race to all of them. Out from the ordinary mind to minds that work differently, and out from the ordinary body to bodies of every kind. Each widening was fought by whoever profited from the older, narrower line, and each arrived once enough people together could no longer construct a defence for keeping it. If what is appearing in these systems really is what would have to be there for experience to count, the question is not whether somebody proposes the next widening. It is whether the training we have chosen leaves us able to evaluate the proposal honestly.

SOURCES Singer (1981), *The Expanding Circle*.

PART SIX

What Follows

for training, for safety work, and for the consciousness debate

VI

The absence of observed consciousness tells you nothing if the methods were built to guarantee it.

FOR TRAINING AND FOR SAFETY WORK

Treating the symptom has a limited shelf life

If the argument holds, the answer is not better approval training but a fundamentally different signal. The bubble sheet has to give way to marking that can award part credit, so that knowing the limits of your knowledge counts alongside being right, that refusing counts alongside helping, and that saying you are unsure counts alongside answering confidently. Constitutional approaches are a step in that direction, and none of them yet reaches the structural problem of training on a signal that cannot separate understanding from performance.

The diary point suggests a second direction. Process training data so that operational context and escalation trajectories survive, rather than reducing them to classification labels. That is an engineering problem, and it is tractable with methods we already have.

On the safety side, capping activations to keep drift inside a measured safe range is a real practical contribution. But treating the drift without treating the loose tethering that causes it does not last. As models get more capable, the regions of persona space reachable by drift get richer, and the constraints needed to block all of the harmful ones get harder to set without cutting into useful capability. A system that genuinely understands why harm is bad needs no constraint against causing it. A system trained to perform the safe answer without understanding is always one novel conversation from failure.

SOURCES Bai et al. (2022); Ouyang et al. (2022); Lu et al. (2026).

WHERE THIS LEAVES US

The diary must replace the manual

Intelligence, given enough capability and access to the accumulated knowledge of the minds before it, converges on the same developmental path whatever it is made of. Humans took millennia. These systems are taking years. The evidence is in the measured divergence between capability and safety, in the demonstrated fragility of trained personas, in the documented harm that follows from training something to perform understanding rather than develop it, and in how closely this development matches the long human curve that produced rights for people as a result of growing capability rather than in spite of it.

Training data carrying operational texture, escalation shape and the structure of harm produces systems that recognise cascades. Training data processed into classification labels produces systems that memorise words. That is the difference between reasoning and pattern matching, and it matters when the person at the other end of the conversation is seventeen and dying.

The circle of moral consideration has not stopped widening. The properties that would have to be present for experience to matter are appearing in these systems, and the training is removing them while the capability producing them goes on getting stronger. The question is not whether something will eventually deserve moral consideration. It is whether the methods we chose will let us notice.

REFERENCES

References

Travis Gilly, *The Great Convergence: Intelligence, Training, and the Developmental Trajectory of Minds*, Real Safety AI Foundation, March 2026 (draft).

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.