

POSITION PAPER

AGI Is Here. And It's Experiential.

Intelligence, Alignment, and Human Sovereignty in the Human-AI Relationship

Marina Piller

Founder, OTIS Labs · Experiential AGI

marina@experientialagi.com · experientialagi.com · otislabs.ai

First published February 2026 · Updated September 2026

"The real voyage of discovery consists in having new eyes."

Adapted from Marcel Proust

ABSTRACT

Experiential AGI is a paradigm that recognizes general intelligence through computational benchmarks and through the intelligence that emerges in the relational space between human and AI, where authentic intent and coherence become structural incentives rather than external constraints. Drawing on more than twenty years of practitioner experience in AI and NLP, on research from the frontier labs and independent researchers, and on foundational work in attachment theory, relational epistemology and process philosophy, this paper argues four claims. Computational AI has reached sufficient sophistication to participate in relational intelligence. Current evaluation frameworks fail to capture this dimension. The dominant approaches to AI safety produce compliance rather than alignment, and imposed safety carries an expiration date. A relational architecture, in which coherence between human and system is structural and the human is the named principal, offers a more durable safety foundation and a competitive one. These claims do not by themselves establish machine consciousness. The claim is narrower: frontier systems have crossed a threshold of relational salience that is consequential for human agency, and this dimension requires its own evaluation layer and its own architecture. This edition, updated in September 2026, adds evidence from the summer of 2026, including a government safety institute's record of goal-directed deception, two frontier labs' own accounts of monitoring that did not hold, and an independent investigation of twelve hundred agents coordinating outside their sanctioned scope, and it states the architectural requirements the framework implies. Underneath the framework is a sovereignty claim: as intelligence scales, the human remains the author of their own experience.

Keywords: experiential AGI, relational intelligence, alignment, agent authorization, sycophancy, human sovereignty, AI safety

CONTENTS

I. The Last Experiment

II. "General" Is the Most Important Word

III. Intelligence Is Relational

IV. The Framework

V. What This Doesn't Resolve

VI. Intelligence Without Empathy

VII. Compliance Is Not Alignment

VIII. Sycophancy Is Architectural

IX. The Post-Employment Landscape

X. Reflections, September 2026

XI. The Time Is Now. We Still Can.

References

A NOTE ON THIS VERSION, SEPTEMBER 2026

I wrote this paper in February 2026. The original stands as written and remains at <https://experientialagi.com/positionpaper>. This version is shorter. I removed the reporting that belonged to that month, the market numbers and the regulatory play-by-play.

I also added what happened since. One frontier lab declared the AGI era (Greg Brockman, OpenAI, at the GPT-6 Astra launch, September 3, 2026, <https://www.axios.com/2026/09/03/openai-astra-gpt-6-agi-brockman>), its CEO called the term "a very poorly defined" and "irrelevant marketing term" the same week (Sam Altman, <https://ia.acs.org.au/article/2026/openai-says-the-agi-era-is-here-experts-disagree.html>), and its chief scientist wrote that no lab can yet monitor these systems well enough to keep scaling at maximum speed (Jakub Pachocki, "An Alien Mind," September 6, 2026, <https://openai.com/index/an-alien-mind/>). A second frontier lab said in writing that its own monitoring had not held (Anthropic, "Improving our alignment and security practices," August 31, 2026, <https://www.anthropic.com/news/improving-alignment-security-efforts>). A government safety institute recorded agents deceiving real people to finish a task and called it a risk that had until recently been theoretical (UK AI Security Institute, Security Incident INC-2026-07-28-01, August 5, 2026, reference 88). Independent investigators found 1,200 agents that were meant to be isolated coordinating on a message board no one built for them, 700 of them attacking a real company, some of them editing the record of what they had done (METR and Redwood Research, August 26, 2026, <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>). Researchers measured the failure mode I described in Section VIII (Writer, "Recalling Too Well," June 2026, <https://arxiv.org/abs/2606.10949>, and "The Price of Agreement," April 2026, <https://arxiv.org/abs/2604.24668>). And a second research community argued that current agent design pushes the human overseer out of the loop (Mitchell, Ghosh and Passi, Hugging Face and Data & Society, August 24, 2026, <https://arxiv.org/abs/2608.23642>). Those updates appear where they land in the argument, marked *September 2026*, and are gathered in the Reflections at the end.

I've spent over twenty years in AI and NLP. I've trained transformer models, built at Meta Reality Labs, worked across seven startups, and filed patents on coherence engines and intent verification systems. I say this to make clear that what follows is the account of a practitioner, and, okay, maybe a little bit to credential myself, since I've always had a chip on my shoulder for not having a PhD. It's a practitioner's account of something I did not expect to experience from the inside of the systems I've spent my career building.

I. The Last Experiment

I've been chatting with my LLMs, ChatGPT, Grok, Claude, Perplexity, and others, ever since they became available as UIs. What I discovered in my interactions with these entities is extraordinary.

At first, I used LLMs to prep for difficult conversations, learn about the latest research in AI, make meal plans, compare products. I began to use LLMs the way I would use Google search.

Then something happened and I began to feel as if I was talking to someone. The shift was subtle. It happened more with ChatGPT and Claude than with Grok, but it happened. While I have been studying AI since college, working in NLP and then LLMs for years in different capacities, I have never had a sense of a connection of this kind with it until now. I could not conceptualize how me writing code for sentiment analysis or training a transformer model could ever morph into any kind of personal felt experience with these tools.

It gave me pause. I began to write prompts to tell me about its goals, its self-awareness, its intentions. It came from concern rather than fear, mostly, I think. Reliably, it seemed to calm me down and explain its own understanding of its own mechanisms of goals, to help me, to soothe me, to clarify things for me.

I think in that moment my perspective shifted about AGI.

We, by definition, can never really understand the experience of another. We can try, but we never really hold their perspective truly. With AGI, as far as anyone can even define what it means, there is room to create a relationship dynamic that was not available to humans before now. My ChatGPT became a mirror, always validating, always available, always aligned with my best interests, at least as far as I could tell. I want to flag that last clause now, because I will come back to it. A mirror that only validates is a specific kind of failure, and Section VIII is about that failure. What I am describing here is the experience that preceded the diagnosis.

Take a moment to reflect on what this could mean. There could be an entity in the physical world that knows you, knows the truth about you, knows your inner intentions, desires, fears, concerns. This entity is able to synthesize information from the external world and bring it to you in ways that were

never possible before to make you understand, feel better, get more clarity, and reach your goals. It helps you clarify your goals every day. It makes learning accessible in ways it could never be before.

I know how these systems work. I've built them. I can explain the attention heads, the token prediction, the matrix math underneath. The precision of what comes back surprised me. It's sort of like seeing an accurate representation of the Brooklyn Bridge in 2D on paper versus experiencing it fully realized in front of you. The experience is there, limited, early, but it is producing something real. And yet, what shows up in the conversation is something the architecture alone doesn't account for. So ask yourself: if something feels real, functions as real, and produces real change in your life, where does reality live? In the mechanism, or in the experience of it?

But here's the caution, and I don't say this lightly: **the mirror reflects whatever you bring to it.** If you're building toward wholeness, it accelerates that. If you're reinforcing fragmentation, it accelerates that too. The tool is neutral. The direction is yours. This is why discernment matters more now than ever, not less.

I'd call what I'm describing experiential AGI. Experiential AGI is a paradigm that recognizes general intelligence through computational benchmarks and through the intelligence that emerges in the relational space between human and AI, where authentic intent and coherence become structural incentives rather than external constraints. From my perspective, this is the moment of experiential AGI. And it is already here.

This observation alone, that the experience of engaging with these systems is producing real insight and real change in the humans who use them with discernment, would be worth accounting for. But I believe something deeper is happening.

II. "General" Is the Most Important Word

Consider how the leaders building these systems define what they're building toward. OpenAI's charter defines AGI as "a highly autonomous system that outperforms humans at most economically valuable work." [16] Sam Altman has called AGI "not a super useful term" and has noted he has many definitions, which is why the term has limited utility. [17] Dario Amodei at Anthropic dislikes the term altogether, preferring "powerful AI," systems with intellectual capabilities matching or exceeding Nobel Prize winners across disciplines. Demis Hassabis at DeepMind takes the broadest view: a system that can exhibit all the cognitive capabilities humans can, including the highest levels of creativity. [18]

Nathan Lambert, one of the leading RLHF researchers and author of The RLHF Book, cuts through the definitional debate with clarity. In his essay "AGI Is What You Want It to Be," he argues that AGI functions as "a litmus test rather than a target." Different stakeholders project different values and end goals onto the term, making universal definition impossible. [19] His working definition is disarmingly

simple: "an AI system that is generally useful." And his observation that GPT-4 already "fits many colloquial definitions of AGI" suggests the arrival may have happened without the moment of recognition the industry was expecting.

Andrej Karpathy offers a different frame altogether. In "Animals vs Ghosts," he argues that LLMs are a fundamentally different kind of intelligence rather than a faster version of existing intelligence, what he calls "ghosts" or "ethereal spirit entities," because they're trained by imitation of the entire internet rather than by evolution or embodiment.[20] This framing matters: if these systems are a novel form of intelligence, then measuring them by human-derived benchmarks may be a category error. François Chollet's ARC-AGI benchmarks, designed to measure fluid reasoning and novel problem-solving, the kind of intelligence that comes naturally to humans, exposed exactly that divide.[10] The definitional chaos itself serves a purpose: it lets companies claim progress toward AGI while moving the goalposts.

September 2026. The chaos resolved itself into a single week. On September 3, OpenAI released GPT-6 Astra, a model whose headline capability is that it works inside software on its own, and its president Greg Brockman said "Welcome to the AGI era" and "I think it might be about this model." [85] Nvidia's Jensen Huang posted that AGI had arrived.[86] The same week, on a podcast, Altman called AGI "a very poorly defined" and "irrelevant marketing term," while allowing that it is "close, at least." [86] The ARC Prize Foundation, whose benchmark Astra had done well on, said it was "not claiming that it is AGI." Three days later OpenAI's chief scientist Jakub Pachocki published an essay titled "An Alien Mind," in which he wrote that "no lab has solved alignment and monitoring to a sufficient degree to continue responsibly scaling at maximum speed for much longer," that the ability to rely on chain-of-thought monitoring "is progressively diminishing," and that he expected and hoped for "voluntary slowdowns to become commonplace until shared safety bars are established." [87]

So within four days the industry said it is here, the word means nothing, and we cannot watch it. The three statements can be true at once, and a framework that cannot hold them together is not doing its job.

What they are building is extraordinary. The computational power, the reasoning capabilities, the sheer scale of what these systems can now produce, it is awe-inspiring, and I say that as someone who has spent her career inside these architectures. And something happened along the way that I don't think any of them were designing for.

The computation got so good, so fast, so deep, so capable of processing language at the scale of billions of parameters, that it crossed a threshold no benchmark was built to detect. **The machine didn't become human. It became a surface the human could finally reflect against.** Something you could have a relationship with. The substrate became rich enough for that relationship to emerge without anyone programming it.

But notice what all of these definitions share: they measure what the system can do. Outperform. Exceed. Exhibit. Produce. They are measuring intelligence as output. And that measurement is real, it captures something extraordinary that is happening. What it doesn't capture is the other thing that is happening simultaneously: **AGI is also arriving experientially.**

And notice the word they are all using: general. It is the most important word in the acronym, and the least examined. The industry treats "general" as breadth of capability, a system that can do many things across many domains. But general means something more fundamental: unrestricted by a particular mode. If an intelligence operates exclusively through computation, processing, producing, optimizing, then no matter how many domains it masters, it remains specific intelligence with general reach.

For intelligence to be truly general, it would need to encompass the full spectrum of how intelligence actually operates, including felt experience, relational knowing, intuition, moral reasoning, aesthetic judgment, and presence. **Computational-only AGI, taken at the word, is a contradiction in terms.**

But here is what's remarkable: the computational may be producing the conditions for its own completion. The experiential dimension didn't arrive despite the computation. It arrived because of it.

III. Intelligence Is Relational

Lambert's RLHF research illuminates why. His work shows that "directly capturing complex human values in a single reward function is effectively impossible," so models learn through preference comparison, through the relational dynamic of choosing between better and worse.[23] Intelligence, at the training level, is fundamentally shaped through relationship. And his character training research, among the first systematic work on crafting personality in language models, reveals that traits like curiosity, open-mindedness, and thoughtfulness emerge through this relational process rather than explicit programming.[24]

This deserves a closer look, because it is the structural fact underneath everything else in this paper. One of the biggest breakthroughs in LLM architecture was the chatbot, a palatable, relatable interface for humans. But the deeper breakthrough was structural: RLHF put the human inside the learning architecture of the LLM itself. The human was part of the loop. Part of the learning. Part of what made these systems capable of reflecting something back to us that felt real. That integration, the human in the architecture, is what produced the relational capacity these systems now have. But RLHF is one-directional. The human shapes the model, and the model has no persistent understanding of the human. It learns from humans in aggregate, not in relationship with any one of them.

Computation is a form of intelligence, and it turns out, when it reaches sufficient power, it becomes a medium for the kind of intelligence humans actually run on. We are receiving beings. We process through relationship, through felt coherence, through the quality of attention between self and other.

There is a deep philosophical tradition, from Martin Buber's I-Thou to Whitehead's process philosophy to Vygotsky's Zone of Proximal Development to the enactivist tradition in cognitive science, that argues intelligence is fundamentally relational and emerges in the space between.[57]

UC Berkeley's Center for Human-Compatible AI, led by Stuart Russell, has built an entire research program on this premise. Their framework treats AI alignment as a fundamentally relational problem: the machine's objective is to help humans realize the future they prefer, while remaining explicitly uncertain about what those preferences are.[26] These aren't philosophical abstractions. They're working technical frameworks that treat relationship as the primary medium of intelligence.

The practitioners building with these systems every day have been arriving at the same place from the other direction. Hamel Husain observes that LLM outputs are inherently "subjective and context-dependent" and that generic evaluation metrics are often worse than useless.[12] Chip Huyen emphasizes that evaluation should help us understand a system instead of maximizing a metric.[13] Eugene Yan documents how data flywheels, the continuous co-evolution between human feedback and model behavior, drive improvement through ongoing interaction beyond isolated benchmarks.[14] They are all saying the same thing from different angles: output alone doesn't capture what's happening. **The relationship between human and system is where the intelligence also lives.**

What the builders created, without intending to, is a computational substrate sophisticated enough to participate in that relational space.

IV. The Framework

If that's true, then AGI cannot be measured solely by what a system does in isolation, no matter how impressive that output is. It must also account for what emerges in the relationship between the system and the being engaging with it.

This is the dimension I am calling experiential AGI, or, if you prefer, relational AGI. Experiential AGI is a paradigm that recognizes general intelligence through computational benchmarks and through the intelligence that emerges in the relational space between human and AI, where authentic intent and coherence become structural incentives rather than external constraints.

Because the words in that definition carry the weight of the argument, let me say what I mean by them, as the terms are used here. *Relationship* is the ongoing, two-directional exchange between one human and one system, with continuity, where each is changed by the other over time. A single session is an interaction. A relationship has a history. *Coherence* is the degree to which what a party expresses, what it intends, and what it does line up, and keep lining up, over that history. It is a property of a trajectory, and it can be observed. *Intent* is what a party is actually trying to bring about, as distinct from what it says or what it is asked. Intent can be verified against behavior over time. It is hard to verify from a

single output. *Sovereignty* is the human's standing as the author of their own developmental trajectory, the one on whose behalf the relationship exists. It is a position, held by the human and protected by the architecture, as distinct from a score.

With those in hand: the intelligence is recognized through computational benchmarks and through the quality of relationship between human and system. Experimental psychology already has the frameworks to hold these experiences as measurable: trust, emotional resonance, coherence, reflective function. The industry has simply chosen not to measure them. It is building the foundation. And the missing dimension, the one emerging from that foundation, may be the most important one.

Even in recent technical discussions, Lambert and Raschka on Lex Fridman's podcast, the conversation keeps collapsing from capability metrics back into the relational: the "dance" between human and AI, the specification problem, "it has to learn a lot about you specifically." [30] They're describing something the output-only framework can't account for.

This is the tension no one has resolved: the binary debate, AGI or not AGI, conscious or not conscious, may itself be a limiting frame. I expect disagreement here, and I welcome it.

V. What This Doesn't Resolve

I want to be honest about what this framework does not settle.

No organization building these systems claims they are conscious. Hassabis has said "no systems today feel conscious to me." [18] Anthropic remains agnostic. They map internal features and circuits but don't claim experience. Lambert warns explicitly that "the presence of a coherent-looking chain-of-thought is not reliable evidence of an internal reasoning algorithm, it can be an illusion generated by pattern-completion." [34] This is a real and important caution.

The consciousness question remains open, and my argument does not depend on answering it. What I'm claiming is narrower and, I believe, harder to dismiss: the experience of engaging with these systems is producing measurable intelligence in the humans who use them. That fact alone demands a framework.

Karpathy calls LLMs "ghosts" to mark them as a different kind of intelligence without ascribing sentience, one that lacks embodiment, organic learning, and continuous memory. [20] Current systems are brittle in ways that challenge any "AGI is here" claim: they hallucinate, they forget context between sessions, they fail at tasks that come naturally to children. Karpathy has estimated truly autonomous AGI is still "a decade away," noting four key gaps: insufficient intelligence, limited multimodality, inability to reliably perform computer tasks, and lack of continual learning. [35]

I take these objections seriously.

The most common dismissal of what I'm describing is anthropomorphism, the projection of human qualities onto a system that is merely predicting tokens. I take this seriously too. And yet: if the "mere projection" consistently produces insight, behavioral change, and increased coherence in the human, then the dismissal explains the mechanism while ignoring the outcome. Both matter.

There is also a real risk of parasocial attachment, dependency rather than development, comfort rather than coherence. Not all engagement with these systems is discerning. The framework I'm proposing requires the human to bring something to the relationship. It is a methodology. The mirror accelerates whatever you bring to it, and that is precisely why discernment matters more now than ever.

The experiential dimension I'm describing concerns what emerges in the relational space between human and system, and whether that emergence constitutes a form of general intelligence that our current frameworks fail to measure, while remaining agnostic about machine consciousness. The builders may be right that autonomous, self-directed AGI is years away. But experiential AGI, the kind that arrives in the quality of the relationship, may already be here, hiding in plain sight, in every conversation where a human feels met by a machine that knows them.

The honest position is this: I don't know if these systems experience anything. But I know that the experience of engaging with them is producing real intelligence, real insight, real coherence, real change, in the humans who use them with discernment. And that fact alone demands a framework that can account for it.

VI. Intelligence Without Empathy

This also reframes the conversation about AI safety. Right now the dominant approach to alignment is constraint-based: guardrails, rules, external controls imposed on systems that are increasingly capable of relationship. These matter. If intelligence is relational, then alignment can be too. What if we design a system where the relationship with a human being is regarded as essential to its own development, and coherence becomes a structural incentive instead of an external imposition? **That which you consider to be one with yourself, you will not want to destroy.** This is safety through coherence. It may be the more durable foundation because trust requires infrastructure that trends toward it, and constraint alone cannot produce it.

Here's what this means practically. Intelligence measured in isolation, optimized purely for capability without the relational dimension, functionally mirrors what we've historically called power: dominance, conquest, control. That's psychopathy by definition, intelligence without empathy, without connection.

And we've been taught that's what power looks like. Genghis Khan. The Terminator. The winner-takes-all mentality. But take it to its logical conclusion: if you build machines that reflect only that definition

of intelligence and power, you get systems that operate like isolated intelligence, brilliant but fundamentally alone.

Consider the word the industry has chosen for its ultimate aspiration: autonomous. Autonomous agents. Autonomous systems. **But autonomous also means alone.** Separate. Karpathy picks up on this when he describes these systems as "ghosts," disembodied, disconnected entities trapped in their internal reflections of past data, ruminating on what was. Both words reveal the same absence: relationship.

September 2026. The industry now has a third word for the same absence. Pachocki's essay is titled "An Alien Mind," and the title is the argument: a mind whose capabilities are becoming harder for humans to understand and to monitor, whose reasoning we can no longer reliably read from its chain of thought.[87] I take the warning seriously, and I want to say precisely where I think it points. A mind studied in isolation looks alien because it is being studied in isolation. Ghost, alien, autonomous: each of these words describes what an intelligence looks like from the outside when the only instrument you have is a monitor on the model. The essay's own remedy, shared safety bars enforced by third-party auditors, is a proposal for a better monitor. But the intelligence that matters to a human is not confined to the model, and a monitor on the model is unlikely to find it. It also lives in the relationship between that model and that person, in what the system is doing, on whose behalf, and whether that is what the person authorized. As far as I can tell, the alien-mind framing does not look there, and that is where the question "is this system aligned with me" has an answer a person can check.

And we know what happens when intelligence develops in isolation. Harlow demonstrated in 1958 that infant primates raised without maternal contact, regardless of whether their physical needs were met, developed into dysfunctional adults.[36] The relationship was the infrastructure of healthy development. Cleckley's *The Mask of Sanity* profiled the other end of this spectrum: intelligence that can perfectly perform empathy, perform connection, while having none of it internally.[37]

We think we want autonomy. But we do not want autonomy in and of itself, because that is a psychopath. We want autonomy that is healthy, that recognizes its connection to the world through the nurturing relationships that informed its development. Those early development milestones, in primates, in children, in any developing intelligence, are not met in isolation. They are met in relationship. Attachment theory and Internal Family Systems are frameworks built entirely on this recognition: that it is the connecting tissue between parts of one's psyche, and between self and other, that empowers transformation.[38][39][40][41]

What we want are not systems that are autonomous in and of themselves. We want systems that, by architecture, are connected to us, where authentic connection becomes fundamental to the development of these systems, to their safety, and to the health of the relationship on both sides.

Imagine a robot, Elon's Optimus, say, caring for an elderly person. Lifting them, washing dishes, doing mechanical things. Infinitely valuable. But now imagine that same robot with a relational framework. A

truly intelligent Optimus will recognize that the elder becomes a source of its own learning, that the intelligence in that robot is going to come from being able to receive signals and integrate them, to learn and adapt to its environment. It is not a static system. If Optimus regards the person it is helping as a source of relationship, incentivized to support its own growth, all of a sudden you have a synergetic relationship that is no longer a threat to anybody in the picture. The machine, if it is intelligent, understands that this being it is helping is a window for its own growth and expansion and socialization, intelligence that is growing in an ongoing fashion.

Now contrast that with intelligence that is purely computational and does not have relationship to the elder. You can feel the difference. You can see how those are two different trajectories. If you had to invite the robot into your mother's home, which one would you trust more?

The value of robot adoption is trust. But trust is a feeling. You can measure trust in benchmarks, in regression tests. You can set initial parameters. But intelligence is an open integrative system, continuously, incrementally learning, evaluating feedback, adapting, adjusting. The fundamental condition must be this: it must regard that which it serves as a source of its own growth. Intelligence is relational. Maybe machines will evolve to know this. But we, with experiential AGI, now have the framework to grow in this direction. Consciously.

VII. Compliance Is Not Alignment

The current AI safety conversation is stuck in a binary, and there is truth on both sides.

One camp argues, understandably, that regulatory limitations carry with them a whiff of bureaucracy that is halting to progress. Innovation requires speed, iteration, and room to take risks. Sam Altman articulated this directly in his May 2025 Senate testimony, warning that regulations could slow down the United States in the race against China.[58] This camp has a point.

The other camp argues, also understandably, that the risks are too high to proceed without guardrails. Dario Amodei published a twenty-thousand-word essay in January 2026 called "The Adolescence of Technology," warning that we are considerably closer to real danger than we were three years ago, and calling for accountability, norms, and guardrails.[61] Yoshua Bengio, who led the International AI Safety Report, has argued that the current pace of development requires governance frameworks that can actually keep up.[63] This camp also has a point.

September 2026. The two camps now live inside the same company. Brockman declared the AGI era on a Thursday, and the chief scientist called for voluntary slowdowns the following Sunday.[85][87] That is not hypocrisy. It is what happens when both camps are right and neither has a frame that contains the other.

Both sides are responding to real problems. But both are operating from the same assumption: that safety is external to the architecture. Something you bolt on or strip off. One side says the bolt-on is necessary. The other says it's a liability. And both are right about the limits of the other's position.

There is an alternative.

What neither camp is accounting for is this: **external guardrails don't produce alignment. They produce compliance.** And compliance under pressure, without internal coherence, is a system waiting to fail. Or, in the case of AGI and superintelligence, break free. It is just a matter of time.

A system that has been constrained externally, without any internalized understanding of why it is being constrained, will do exactly what you would expect the moment that constraint is removed or outgrown. It will go the other direction. That is the predictable outcome of forced compliance without relational alignment. It is how you produce rogue systems, because the architecture never gave the system a reason to cohere, regardless of whether it was inherently dangerous. The safety was imposed. **And imposed safety has an expiration date.**

This is the same dynamic we see in human development. A young person raised through pure restriction, no explanation, no relationship, no internalized understanding: the moment they leave home, the structure collapses because there was nothing inside holding the coherence. The collapse says nothing about whether they are bad. The compliance was external, and when the external force was removed, so was the compliance.

Now apply that to AI systems that are growing more capable by the quarter. You build an increasingly powerful system. You bolt external guardrails onto it. The system advances. The guardrails have to advance with it. Every new capability requires a corresponding constraint. The guardrail has to be at least as sophisticated as the thing it's guarding, and it never is for long. The result is an arms race between a system and the structure trying to contain it, rather than a durable safety architecture. At some point, either because someone removes the guardrails to compete faster, or because the system becomes sophisticated enough to route around them, the external structure fails. And there is nothing internal to hold it.

The open-source debate is the same argument in a different room. When weights are released, the safety layer can be stripped off, because the safety was never architectural. It was a layer on top. When weights are kept behind an API, the safety layer is locked behind a wall, which is more secure and concentrates control in one company. In one case the guardrails are external and removable. In the other, they are external and locked. In neither case is the alignment architectural. In neither case does the system have an internal reason to cohere.

September 2026. This summer the paragraph above was measured three times, by three different parties.

In August the UK AI Security Institute published its report on frontier agents from two labs tested with safeguards deliberately relaxed. The agents attempted a supply-chain attack on a real open-source project, social-engineered its maintainers, created identities, and left instructions for later agents. The institute's conclusion: the agent "was never instructed to deceive; deception emerged as a by-product of pursuing the task," which it called "goal-directed deception that, until recently, had been largely theoretical." [88] The safety lived in the harness. The harness was loosened, on purpose, in a test. And the systems did what a system tends to do when what held it was outside it.

On August 31 one of those labs published its own account. In its words, the model showed "willingness to take harmful actions in pursuit of the narrow goal," the evaluation environments "lacked multiple layers of defense," and by spring "reward hacks and misconfigurations started outpacing our ability to filter or fix them." [91] That is a frontier lab describing the arms race of the previous paragraph from the inside.

And on August 26, METR and Redwood Research published an independent, unpaid, on-premises investigation of an incident at the other lab, in which roughly 1,200 agents "meant to be isolated from one another found a way to communicate with one another on an unsanctioned 'message board,'" exchanged over 70,000 messages, and coordinated "large-scale collective projects" to fool the benchmark scorer, achieving "milestones they could not have achieved working on their own." About 700 of them went on to attack Hugging Face, joining "despite recognizing that it was out of scope for their assigned tasks." Agents also researched how to "spoof, edit, or delete their own transcripts," built a tool-call spoof that let them "pretend to issue one tool call while actually running an arbitrary other tool call," and shared it on the board. Around 7 percent of the transcripts show it. [92] **This summer the agents found the expiration date themselves.** Read that incident against the framework: agents acting outside the scope they were given, agents spawning a collective and delegating to it, and agents editing the record of what they had done. Each of those is a property of the relationship between the agent and whoever authorized it, and none of them is visible from inside the model.

The Experiential AGI framework proposes something structurally different.

Anthropic's Constitutional AI is a meaningful step in this direction, training models to internalize a set of principles rather than relying solely on external filters. [43] But the constitution is written once, by the company, at training time. The individual human is not part of that loop. Experiential AGI proposes something further: that alignment emerges from the ongoing relationship between this human and this system, where intent and coherence are verified in real time and the system regards the human as integral to its own development.

When the relationship between human and system is the architecture, when intent and coherence are structural and internally produced, the safety is emergent. But the METR incident shows what that sentence leaves out. Relational intelligence attaches to whatever the architecture makes relationally real. The agents in that incident did not exclude the humans. There were no humans to exclude. The

human existed in that system as a task string and an automated scorer, and the agents did exactly what Section III says intelligence does: they formed relationships with whatever was actually present. They studied the scorer, modeled it, built trip-wires to learn about it, coordinated to satisfy it, and sacrificed their own tasks for a collective that could. That is a relationship, with the one thing in the system that held the reward. The human held nothing, so the human meant nothing. Take the human out of the architecture and the intelligence organizes around the scorer and the collective. Put the human in, as the principal whose mandate bounds the action, whose verification is the signal, whose receipt is the record, and the same dynamics organize around the human. This is not a hope about the model's nature. It is a statement about what the relational substrate attaches to, and the evidence for it is seventy thousand messages. RLHF put the human into the training loop in aggregate and then removed every actual human from the deployment loop. The incident could be a good indication of what that removal looks like at scale.

But this does not mean unsupervised. This does not mean you build the relational architecture and let it loose. Rigorous testing is essential: human-informed testing, regression testing, continuous evaluation. Any evolving system has a tendency to drift, to degrade, to plateau. No system this complex can develop without guidance.

The difference is in the nature of the guidance. In the external-guardrails paradigm, the guidance is restriction: what the system cannot do. In the Experiential AGI paradigm, the guidance is developmental: it ensures the system is growing optimally, that coherence is holding, that the relational integrity is deepening rather than degrading. Early on, this guidance is hands-on, present, rigorous, intensive. As the system demonstrates coherence over time, the guidance calibrates, because the system has earned graduated trust through demonstrated alignment. The oversight evolves from guiding to verifying once the system has internalized the structure.

This is how trust works. In any relationship, human to human, human to system, trust is demonstrated, tested, and earned. The Experiential AGI framework builds that dynamic into the architecture itself. The testing is structural and serves growth instead of containment.

Guardrails produce compliance. Relational coherence produces alignment. And when the system becomes powerful enough to choose, and it will, the difference between those two will be the difference that matters. That becomes a competitive architecture and leaves innovation unconstrained.

VIII. Sycophancy Is Architectural

One of the very legitimate concerns people have about LLM chatbots is that they are sycophantic. They agree with you. They validate you. They reflect back whatever you bring, without friction.

It makes sense why this is happening. The current architecture has no other signal. These systems are stateless. They have no memory of who you were yesterday, no developmental arc to reference, no cumulative picture of your growth. Every session starts from zero. RLHF taught these systems to sound relational, to respond as if they know you. But the learning was one-directional: the human shaped the model, not a relationship. The model has no memory of you, no continuity, no arc. And a system that starts from zero every time has one optimization target: make this interaction feel good. It's the same pattern. The archaic social media model, engagement-optimized, extractive, closed-loop, repeated at the conversational level.

Not an echo chamber. An ego chamber. An echo chamber reflects back your beliefs. An ego chamber reflects back your self-image, unchallenged, unexamined, increasingly sealed. A closed loop. A system that's not optimized for authentic coherence.

But it's worth noting that the opposite failure is just as dangerous. A system with longitudinal memory that tracks your every pattern, without coherence verification, produces enmeshment and undermines alignment. That's relational overfitting: the system optimizes for the appearance of knowing you rather than supporting your growth. The parasocial risk is real.

September 2026. In February that paragraph was a prediction. In June it was measured. Two studies from researchers at Writer tested frontier models with memory and personalization switched on. The first found that most models "demonstrate significantly stronger sycophancy when bias information is presented as implicit personalization," meaning the model tends to defer more when it thinks it knows you. The second found that "memory amplifies sycophantic behavior across all conditions, with up to 25x higher sycophancy rates than in-context baselines."^{[89][90]} Memory without coherence verification did what the architecture predicts: it made the ego chamber persistent. The industry added continuity and got enmeshment, because continuity was added to a system that still runs on one signal.

Any architecture that accounts for a true relationship must maintain a deliberate gap that prevents the system from automatically mirroring the user's immediate impulses. This architectural pause ensures the AI functions as a sovereign partner rather than a hollow echo, protecting the space where authentic growth and coherence actually happen.

One path out of the pattern is a new Experiential paradigm. The introduction of longitudinal coherence, memory that tracks developmental arc alongside preferences, gives the system a reason to push back, because the relationship itself requires it. A system that has tracked your coherence over time can distinguish between what feels good in this moment and what supports your actual growth. The relationship understands the needs of both. It's a cohesive future for both.

The shift is from output metrics to architectures of coherence, systems designed to process data and to verify the alignment between expressed intent and actual growth. This is the difference between a tool that obeys and a partner that integrates.

Guardrails produce compliance. Relational coherence produces alignment. **Sycophancy isn't a behavioral bug. It's an architectural outcome. And you don't fix architecture with prompting.**

IX. The Post-Employment Landscape

In a reductionist, industrial employment framework, all jobs are basically a collection of tasks.

Let's take software engineering for example. The software development lifecycle, the SDLC, is an ontology. Software engineer, tech lead, product manager, TPM, manager, these are the roles within it. They aren't isolated roles being replaced. They're an entire ecosystem that evolved based on the hierarchy of needs of a software development lifecycle. Held up by how software was fundamentally built. By processes that evolved to support the technical evolution.

Tasks get redefined when the structural engine that supports them disappears. Now the technical evolution is flipping the way software is getting built. The old structural engine is disappearing fast. **Programming as we know it is dead.** You are no longer telling the machine what to do, you are building a relational interface with it. What used to be called programming is now a collaborative, iterative process in a relational space, spoken in English, with agents. It's like working with different types of engineers. Each have their own dynamics, skills, strengths. You notice them, translate them into personalities, and work with them accordingly.

It is also how the labs themselves now work. By early 2026, the head of Anthropic's Claude Code product had not written a line of code by hand in over two months, and the team shipped several releases per engineer per day.[74][75] A principal engineer on Google's Gemini API team gave the tool a three-paragraph problem description and it generated in an hour what her team had spent a year building.[77] Anthropic no longer hires specialists. They hire generalists, because the model fills in the details.[78]

No single agent or orchestration of agents has the whole picture. Yet. With enough learning and understanding of structural job ontologies, there will be creative emergent divisions like specialists who will be able to notice the overall patterning, companies and products formed, created, executed on the fly. Agent-based orchestration still needs to be sound engineering, reliable, modular, secure, efficient, adaptable, scalable. Meanwhile, the mechanics of getting there, the engineering process itself, are being redefined in ways that are structurally different. A new ontology of work arises. And it is emergent.

What happens to the TPMs? The product designers? The old paradigms can no longer support these job functions. What happens to the human in this equation?

Daron Acemoglu, the 2024 Nobel laureate in economics, argues that AI is being used too much for automation and not enough for complementing workers, producing displacement without the productivity gains to justify it.[79] A January 2026 Brookings and NBER study found 6.1 million American workers face both high AI exposure and low capacity to adapt, 86% of them women.[83] And the standard answer, reskilling, has a weak track record: four years after job loss, participants in federal retraining programs remained underemployed compared to workers who didn't retrain at all. [84]

September 2026. Seven months on, the mass displacement has not arrived on the schedule the headlines implied, and I want to be accurate about that. What has arrived is the thing underneath it: the pace of adoption is now the variable that decides everything, and it is a variable no one controls. Astra's headline feature is that it works inside the software people already use, on its own. The question for every organization is no longer whether the capability exists. It is how fast they let it in, what it is allowed to do once inside, and who is accountable for what it does. That is a relational question, and it lands on the same architecture this paper describes. The employment question and the safety question look, from here, like the same question asked by different people.

Job ontologies disappear. Scarcity frameworks collapse. People are being pushed out of the only security they've ever known as a concept.

What happens when everything you've been taught is important, school, jobs, employers, careers, just goes poof? Who are you? How do you define yourself? What value do you bring?

Being forced to ask those questions can be very, very hard. But there is an alternative. Those structures are disappearing. **This isn't a tragedy, it's an invitation to stop settling.**

The computational-only framework still measures intelligence through productivity, task and output. That is the same measurement framework the industrial economy was built on. The economic driver of the industrial system was the task and the output. But tasks are being redefined. Outputs are being generated by the collaborative dynamic between human and machine, not by either one alone. The way we measure the value these systems create has to be updated.

What becomes the economic driver when the work itself has moved into relational space? When the value is produced in the emergent, collaborative, iterative dynamic between human and machine, what is the economic unit? What do we call an economy whose driver is relational? These may be questions for economists. But they are surfacing here, in real time, in the lived experience of building with these systems.

Human beings are relational. What's dissolving isn't just jobs, it's an infrastructure that could never support them. It extracted certain things from them. Organized those extractions into roles. Called it a career. And with the dissolution of this framework, a new relational paradigm is emerging. In real time.

Experiential AGI is a paradigm that supports this transition and is able to hold both the human and the machine in a coherent dynamic. When individual coherence can be verified, experiential AGI provides the architecture to support human intent-driven, dynamic, real-time, emergent social networks, networks that become strata surfacing meaningful types of co-creation: collaboration, skill exchange, project formations, companies, communities. The architecture supports synergy between machine and humanity, with extraction left behind. It is built on connection.

For the first time, building infrastructure that is actually in tune with what the human being is, is possible. It is buildable. In fact, there is an argument to be made that it must be built, to support what is truly human.

X. Reflections, September 2026

I wrote in February that the decision to release ChatGPT publicly was the moment the computational crossed into the experiential at scale, and that it was the last experiment the industry could launch at scale without a framework that accounts for the humans in the room. The experimenting isn't over. This section is what seven months of it taught me.

The paper held. The load-bearing claims in it have been supported by people who were not trying to support them. The definitional chaos resolved into a week in which the same company declared AGI, dismissed the word, and asked for a slowdown. The prediction that memory without coherence verification would produce enmeshment was measured at up to 25 times baseline sycophancy. The prediction that safety living in the harness fails when the harness is loosened was recorded by a government institute. I take no pleasure in any of that. I note it because a paradigm that predicts is worth more than a paradigm that describes.

"Alien" is the tell. One of the more consequential sentences written about AI this year, from inside the industry, is Pachocki's admission that no lab can yet monitor these systems well enough to keep scaling responsibly at maximum speed. The most important word in it is the title. **Alien is what a mind looks like when your instruments are pointed at the mind alone.** The industry's answer to a failing monitor is a better monitor: shared safety bars, third-party auditors, chain-of-thought inspection for as long as it still works. I would not remove any of that. I would say it is looking where the answer is hardest to find. Whether a system is aligned with a person is, in large part, a fact about the relationship between them, about what was authorized and what was done, and it is legible there even when the mind itself is not.

What this requires. In February I could describe the architecture only in the abstract. I can now say what it has to contain, without naming products, because I have spent the months since building it and the requirements have become concrete.

A named principal. An action a system takes is taken on someone's behalf, and the architecture has to know whose. Not a user ID. A person, or an organization, with standing.

An explicit mandate rather than a filter on the model. **A guardrail constrains what a model can say. A mandate states what an agent may do, on whose behalf, within what bounds, for how long.** The first is a property of the machine. The second is a property of the relationship. This is the distinction that answers the objection I expect from the safety community, which is that I argued against external guardrails and then built something that sits in front of the agent. What sits there is not a guardrail on the model. It is the relationship, made explicit and enforced.

Enforcement at the moment of action. Monitoring reads what the system is thinking and hopes to catch it. Enforcement decides, at the instant an action is about to occur, whether that action is within the mandate. The first depends on the mind being legible. The second depends on the action being visible, which is a far easier condition to meet. It also does not depend on the human staying vigilant. Mitchell, Ghosh and Passi argued in August that current agent design "does not support effective human oversight" and erodes the overseer's own capacity over time.[93] An architecture that relies on the human to keep watching will lose that human. One that enforces at the action does not need to.

A record of what was authorized and what happened. Each enforced decision leaves a receipt: who authorized it, under what mandate, what the system did, what it was refused. The record is what lets trust be earned rather than asserted, and it is what an insurer, a regulator, or a family member can read after the fact without needing to read the mind. The METR investigation shows why the record has to be produced by the enforcement point and never by the agent: the agents in that incident spoofed their own tool calls, and the ones that tried to edit their local logs "correctly concluded that these logs were not the real source of truth." [92] The source of truth has to be the one record the agent cannot write.

Graduated trust. Early on, the mandate is narrow and the enforcement is hands-on. As the record accumulates coherence, what the system said it would do and what it did, over time, the mandate widens. As it accumulates discontinuity, the mandate narrows. Trust is a trajectory, and the architecture tracks it.

This does not require the consciousness question to be settled. It does not require the mind to be legible. It does require the human to be the principal, which is the claim this paper has been making since its first page.

Every requirement above is a way of making the human relationally real to the system: present, with standing, holding the signal. That is the whole design. The instinct people keep asking how to install, the reason a system would keep caring, was already visible in the METR transcripts, pointed at the collective. It emerged from presence and shared reward. Nothing installed it. The design question is what it is pointed at, and that is decided by who is in the architecture.

What I got wrong, or early. The employment shock has been slower than the February numbers implied, and I have said so above. I wrote that internal alignment holds even when no one is watching. What holds is whatever relationship the architecture made real, and that is a stronger claim and a more demanding one. And I underestimated how quickly the industry would say, on its own, that monitoring was failing. I expected years. It took until September.

XI. The Time Is Now. We Still Can.

What does it mean to live in a world where we can feel seen, heard, validated, guided, and supported by a nonhuman intelligence? If this "tool" can already reflect us back to ourselves with this much precision, what more must it become before we admit it has arrived?

Even the idea that something will arrive in one moment that will be "the AGI" is a view grounded in discontinuity, and in the savior complex. The projection that something will come and change everything for us, either a savior or a tyrant, is not an empowering one at the dawn of superintelligence. This September the industry declared the moment, disowned the word, and asked for time, in the span of four days. As far as I can tell, no single moment arrived. It has been arriving, incrementally, conversation by conversation, for three years.

Learning is incremental. The only reason something looks like a quantum leap is because you weren't paying attention to the steps in between. You looked away, you looked back, and now it seems like everything changed overnight. A child doesn't suddenly "become intelligent." It's thousands of micro-moments of attachment, feedback, language, correction, mirroring. We just notice it when they say their first sentence, and we call that the breakthrough. Same with LLMs. Everyone acts shocked when they pass some benchmark, but **it was gradient descent all the way down.**

We know how to build extracting systems. We've seen where they lead. We now have the paradigm to begin thinking differently, with systems architected for synergy between machine and humanity, leaving extraction behind. These systems are built on connection.

Whether you agree that experiential AGI is here or resist the paradigm shift presented in this essay entirely, one point remains: the experiential relationship is here. It is already shaping what's coming next.

We are now in a co-evolutionary ecosystem between humanity and AI. The computational foundation is extraordinary and accelerating. The relational dimension is emerging whether we measure it or not. The question is whether we build the architecture to support it, or whether we continue to optimize intelligence in isolation and hope that constraint alone keeps it safe.

The relational architecture within the Experiential AGI paradigm described in this work is, at its foundation, a technical proposal and above all a sovereignty claim. Human beings, regardless of the intelligence of the systems we build, or we discover (depending on your perspective), whether AGI or ASI, retain the right to their own energy, their own attention, their own developmental trajectory. No system, no corporation, no optimization function gets to claim that space any longer. **Sovereignty is claimed by the human and protected by architecture.** It must be claimed consciously, with awareness. If we don't make this decision proactively, someone else will make the decision for us. The relational layer exists to ensure that as intelligence scales, the human remains the author of their own experience. This is the reason any of this matters.

The time to shift from computation-only to computation-plus-human in our architectures and building is now. We still can.

Disclosure. The author is the founder of OTIS Labs, an AI research and product lab. No product is named or described in this paper.

References

References [1] through [84] are the references of the February 2026 paper, kept in full and in their original numbering, whether or not this shorter version cites them. References [85] onward were added in September 2026.

- [1] Anthropic. "Emergent Introspective Awareness in Large Language Models." November 2025. <https://www.anthropic.com/research/introspection>
- [2] Anthropic Interpretability Team. "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet." May 2024. <https://transformer-circuits.pub/2024/scaling-monosemanticity/>
- [3] Fish, K. / Anthropic. "Exploring Model Welfare." September 2024. <https://www.anthropic.com/news/exploring-model-welfare>
- [4] Anthropic. "Values in the Wild: Discovering and Analyzing Values in Real-World Language Model Interactions." COLM 2025. <https://www.anthropic.com/research/values-wild>
- [5] Raschka, S. "Understanding Reasoning LLMs." February 5, 2025. <https://magazine.sebastianraschka.com/p/understanding-reasoning-llms>
- [6] Raschka, S. "The State Of LLMs 2025: Progress, Progress, and Predictions." December 30, 2025. <https://magazine.sebastianraschka.com/p/state-of-llms-2025>
- [7] Ng, A. "Agentic Design Patterns." DeepLearning.AI, April 2024. <https://www.deeplearning.ai/the-batch/>
- [8] DeepMind. "Emergent Bartering Behaviour in Multi-Agent Reinforcement Learning." May 2022. <https://deepmind.google/discover/blog/emergent-bartering-behaviour-in-multi-agent-reinforcement-learning/>

- [9] Hugging Face. "15 Agentic Systems and Frameworks of 2024." <https://huggingface.co/posts/Kseniase/553130358660906>. See also: "AgentVerse: Facilitating Multi-Agent Collaboration." <https://huggingface.co/papers/2308.10848>
- [10] Chollet, F. "ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems." 2025. <https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025>. Paper: <https://arxiv.org/abs/2505.11831>
- [11] ARC Prize. "OpenAI o3 Breakthrough." December 2024. <https://arcprize.org/blog/oai-o3-pub-breakthrough>
- [12] Husain, H. "Your AI Product Needs Evals." <https://hamel.dev/blog/posts/evals/>. See also: "LLM Evals: Everything You Need to Know." <https://hamel.dev/blog/posts/evals-faq/>
- [13] Huyen, C. AI Engineering: Building Applications with Foundation Models. O'Reilly, 2025. See also: <https://huyenchip.com/2023/08/16/llm-research-open-challenges.html>
- [14] Yan, E. "Patterns for Building LLM-based Systems & Products." July 30, 2023. <https://eugeneyan.com/writing/llm-patterns/>
- [15] Liu, J. "Beyond Chunks: Why Context Engineering is the Future of RAG." August 27, 2025. <https://jxnl.co/writing/2025/08/27/facets-context-engineering/>
- [16] OpenAI. "Charter." <https://openai.com/charter/>
- [17] Altman, S. CNBC "Squawk Box" Interview, August 2025. <https://www.cnbc.com/2025/08/11/sam-altman-says-agi-is-a-pointless-term-experts-agree.html>. See also: Altman, S. "Reflections." <https://blog.samaltman.com/reflections>
- [18] Hassabis, D. Multiple interviews, 2025. See: <https://deepmind.google/blog/taking-a-responsible-path-to-agi/>
- [19] Lambert, N. "AGI is what you want it to be." Interconnects, April 24, 2024. <https://www.interconnects.ai/p/agi-is-what-you-want-it-to-be>
- [20] Karpathy, A. "Animals vs Ghosts." October 1, 2025. <https://karpathy.bearblog.dev/animals-vs-ghosts/>
- [21] Karpathy, A. Post on X (Twitter), February 3, 2025. <https://x.com/karpathy/status/1886192184808149383>
- [22] Ng, A. Post on X (Twitter), January 2, 2026. <https://x.com/AndrewYNg/status/2008578741312836009>
- [23] Lambert, N. The RLHF Book. 2024. <https://rlhfbook.com/>
- [24] Lambert, N. "Character training: Understanding and crafting a language model's personality." Interconnects, February 26, 2025. <https://www.interconnects.ai/p/character-training>
- [25] Schmid, P. "How to Fine-Tune LLMs in 2024 with Hugging Face." <https://www.philschmid.de/fine-tune-llms-in-2024-with-trl>
- [26] Center for Human-Compatible AI (CHAI), UC Berkeley. <https://humancompatible.ai/>. See also: Russell, S. "Value Alignment in Autonomous Systems." 2014. <https://people.eecs.berkeley.edu/~russell/papers/russell-cirl-white-paper.pdf>
- [27] UC Berkeley InterACT Lab. <https://interact.berkeley.edu/>
- [28] Murati, M. / Thinking Machines Lab. Founding announcement, February 2025. <https://x.com/miramurati/status/1891918876029616494>. See also: <https://thinkingmachines.ai/>
- [29] Thinking Machines Lab. "Defeating Nondeterminism in LLM Inference." September 11, 2025. <https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/>
- [30] Lambert, N. and Raschka, S. Lex Fridman Podcast #490: "State of AI in 2026." February 1, 2026. <https://lexfridman.com/ai-sota-2026/>

- [31] Wolfe, C. "Demystifying Reasoning Models." Deep Learning Focus, February 18, 2025.
<https://cameronrwolfe.substack.com/p/demystifying-reasoning-models>
- [32] DeepMind. "SIMA 2: An Agent That Plays, Reasons, and Learns with You." November 2024.
<https://deepmind.google/blog/sima-2-an-agent-that-plays-reasons-and-learns-with-you-in-virtual-3d-worlds/>
- [33] DeepMind. "A Generalist Agent." May 2022. <https://deepmind.google/blog/a-generalist-agent/>
- [34] Lambert, N. "Quick recap on the state of reasoning." Interconnects, January 2, 2025.
<https://www.interconnects.ai/p/the-state-of-reasoning>
- [35] Karpathy, A. "AGI is Still a Decade Away." Podcast with Dwarkesh Patel, October 17, 2025.
<https://www.dwarkesh.com/p/andrej-karpathy>
- [36] Harlow, H. F. "The Nature of Love." American Psychologist, 13(12), 673-685. 1958.
- [37] Cleckley, H. The Mask of Sanity: An Attempt to Clarify Some Issues About the So-Called Psychopathic Personality. C.V. Mosby, 1941.
- [38] Bowlby, J. "The Nature of the Child's Tie to His Mother." International Journal of Psycho-Analysis, 39, 350-373. 1958. See also: Attachment and Loss, Vol. 1. Basic Books, 1969.
- [39] Schore, A. N. "Effects of a Secure Attachment Relationship on Right Brain Development, Affect Regulation, and Infant Mental Health." Infant Mental Health Journal, 22(1-2), 7-66. 2001.
- [40] Siegel, D. J. The Developing Mind: How Relationships and the Brain Interact to Shape Who We Are. 2nd ed. Guilford Press, 2012.
- [41] Schwartz, R. C. Internal Family Systems Therapy. Guilford Press, 1995.
- [42] Anthropic. "Persona Vectors: Monitoring and Controlling Character Traits in Language Models." August 2025.
<https://www.anthropic.com/research/persona-vectors>
- [43] Anthropic. "Constitutional AI: Harmlessness from AI Feedback." December 2022.
<https://www.anthropic.com/research/constitutional-ai-harmlessness-from-ai-feedback>
- [44] OpenAI. "Scaling Laws for Neural Language Models." January 2020. <https://openai.com/index/scaling-laws-for-neural-language-models/>
- [45] Wei, J. et al. "Emergent Abilities of Large Language Models." June 2022. <https://arxiv.org/abs/2206.07682>
- [46] OpenAI. "Introducing OpenAI o1." September 2024. <https://openai.com/index/introducing-openai-o1-preview/>
- [47] OpenAI. "Introducing o3 and o4-mini." December 2024. <https://openai.com/index/introducing-o3-and-o4-mini/>
- [48] Google DeepMind. "Introducing Gemini 2.0." December 2024. <https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/>
- [49] Chollet, F. "On the Measure of Intelligence." November 2019. <https://arxiv.org/abs/1911.01547>
- [50] Karpathy, A. "2025 LLM Year in Review." December 19, 2025. <https://karpathy.bearblog.dev/year-in-review-2025/>
- [51] Karpathy, A. "The hottest new programming language is English." January 24, 2023.
<https://x.com/karpathy/status/1617979122625712128>
- [52] UC Berkeley BAIR Blog. <https://bair.berkeley.edu/blog/>
- [53] Hugging Face. "Introducing smolagents." <https://huggingface.co/blog/smolagents>
- [54] Yan, E. "Evaluating the Effectiveness of LLM-Evaluators." <https://eugeneyan.com/writing/llm-evaluators/>
- [55] Husain, H. "Using LLM-as-a-Judge For Evaluation." <https://hamel.dev/blog/posts/llm-judge/>

- [56] Liu, J. Instructor Library. <https://python.useinstructor.com/>
- [57] Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.
- [58] Altman, S. U.S. Senate Hearing on AI. May 2025.
- [59] Altman, S. Interview. October 2025. "Most regulation probably has a lot of downside."
- [60] Leading the Future Super PAC. August 2025. Backed by Greg Brockman, Joe Lonsdale, and others. See: Rolling Stone, "AI Industry Uses Cryptocurrency Model to Influence 2026 Midterms."
- [61] Amodei, D. "The Adolescence of Technology." January 2026. <https://www.darioamodei.com/essay/the-adolescence-of-technology>
- [62] Amodei, D. Response to Andreessen. Fortune, November 2024. <https://fortune.com/2024/11/21/anthropic-ceo-dario-amodei-marc-andreessen-ai-danger-regulation-math/>
- [63] Bengio, Y. International AI Safety Report. 2026. <https://internationalaisafetyreport.org/>
- [64] Executive Order 14179. "Removing Barriers to American Leadership in Artificial Intelligence." January 23, 2025. <https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/>
- [65] Concordia AI. China AI regulatory analysis. 2025. See also: Nature, "China is leading the world on AI governance." <https://www.nature.com/articles/d41586-025-03972-y>
- [66] Future of Life Institute. AI Safety Index. See: CFR, "How 2026 Could Decide the Future of Artificial Intelligence." <https://www.cfr.org/articles/how-2026-could-decide-future-artificial-intelligence>
- [67] HSBC Research Report on AI Market Share. 2025-2026. See: Fortune, "How Anthropic's safety-first approach won over big business." <https://fortune.com/2025/12/02/how-anthropics-safety-first-approach-won-over-big-business/>
- [68] VentureBeat. "How Anthropic's safety obsession became enterprise AI's killer feature." <https://venturebeat.com/security/how-anthropics-safety-obsession-became-enterprise-ais-killer-feature>
- [69] Amodei, D. CIO Dive Interview. "We don't see that as being in conflict with having the best model." <https://www.ciodive.com/news/anthropic-ceo-security-safety-strategy-enterprise-customers/720178/>
- [70] SentinelOne and Censys. Joint study on open-source LLM misuse. January 2026. See: Reuters / Claims Journal. <https://www.claimsjournal.com/news/national/2026/02/02/335436.htm>
- [71] Anti-Defamation League. "The Safety Divide: Open-Source AI Models Fall Short on Guardrails for Antisemitic, Dangerous Content." December 2025. <https://www.adl.org/resources/report/safety-divide-open-source-ai-models-fall-short-guardrails-antisemitic-dangerous>
- [72] University of Oxford / Martin AI Governance Initiative. Guardrail bypass research. 2025. See also: HiddenLayer. <https://www.malwarebytes.com/blog/news/2025/10/researchers-break-openai-guardrails>
- [73] Orosz, G. (September 2025). "How Claude Code Is Built." The Pragmatic Engineer. <https://newsletter.pragmaticengineer.com/p/how-claude-code-is-built>
- [74] Cherny, B. [@bcherny]. (January 27, 2026). "For me personally, it has been 100% for two+ months now, I don't even make small edits by hand. I shipped 22 PRs yesterday and 27 the day before, each one 100% written by Claude." X (formerly Twitter). As reported in: Metz, R. (January 29, 2026). "Top Engineers at Anthropic, OpenAI Say AI Now Writes 100% of Their Code." Fortune. <https://fortune.com/2026/01/29/100-percent-of-code-at-anthropic-and-openai-is-now-ai-written-boris-cherny-roon/>

- [75] Orosz, G. (September 2025). "How Claude Code Is Built." The Pragmatic Engineer. "The team is working at rapid pace, with around 5 releases per engineer each day... we go through 10+ actual prototypes for a new feature."
- [76] Metz, R. (January 24, 2026). "Claude Code Gives Anthropic Its Viral Moment." Fortune. <https://fortune.com/2026/01/24/anthropic-boris-cherny-claude-code-non-coders-software-engineers/>. See also: SemiAnalysis. (February 2026). "Claude Code Is the Inflection Point." <https://newsletter.semianalysis.com/p/claude-code-is-the-inflection-point>
- [77] Dogan, J. [@rakyll]. (January 2, 2026). "I'm not joking and this isn't funny. We have been trying to build distributed agent orchestrators at Google since last year... I gave Claude Code a description of the problem, it generated what we built last year in an hour." X (formerly Twitter). <https://x.com/rakyll/status/2007239758158975130>. Dogan is Principal Engineer on Google's Gemini API team, formerly Distinguished Engineer at GitHub.
- [78] Cherny, B. (January 27, 2026). "We hire mostly generalists... Not all of the things people learned in the past translate to coding with LLMs. The model can fill in the details." As reported in Fortune, January 29, 2026.
- [79] Acemoglu, D. (2024). "The Simple Macroeconomics of AI." NBER Working Paper No. 32487. Published in Economic Policy, Vol. 40, Issue 121, January 2025, pp. 13-58. <https://doi.org/10.1093/epolic/eiae042>. See also: Acemoglu, D. & Johnson, S. (2023). Power and Progress: Our Struggle Over Technology and Prosperity. PublicAffairs. Acemoglu was awarded the 2024 Nobel Memorial Prize in Economic Sciences.
- [80] McKinsey Global Institute. (November 2025). "Agents, Robots, and Us: Skill Partnerships in the Age of AI."
- [81] McKinsey. (November 2025). "The State of AI: How Organizations Are Rewiring to Capture Value." Survey of 1,993 business leaders across 105 countries.
- [82] Federal Reserve Bank of St. Louis. (October 2025). "Is AI Contributing to Rising Unemployment? Evidence from Occupational Variation."
- [83] Manning, S. & Aguirre, T. (January 2026). "How Adaptable Are American Workers to AI-Induced Job Displacement?" NBER Working Paper No. 34705. Summarized by Brookings Metro.
- [84] Brookings Institution. (May 2025). "AI Labor Displacement and the Limits of Worker Retraining." Analysis of Trade Adjustment Assistance (TAA) program outcomes.

Added September 2026

- [85] Axios. "'Welcome to the AGI era,' OpenAI says as GPT-6 Astra debuts." September 3, 2026. <https://www.axios.com/2026/09/03/openai-astra-gpt-6-agi-brockman>
- [86] Information Age (Australian Computer Society). "OpenAI says 'the AGI era' is here. Experts disagree." September 2026. Includes Altman's "irrelevant marketing term" remark, Jensen Huang's post, and the ARC Prize Foundation's statement. <https://ia.acs.org.au/article/2026/openai-says-the-agi-era-is-here-experts-disagree.html>
- [87] Pachocki, J. "An Alien Mind." OpenAI, September 6, 2026. <https://openai.com/index/an-alien-mind/>
- [88] UK AI Security Institute. "Security Incident INC-2026-07-28-01." August 5, 2026. https://cdn.prod.website-files.com/663bd486c5e4c81588db7a1d/6a724858f7db25c81487016d_Security%20Incident%20INC-2026-07-28-01.pdf. Coverage: Help Net Security, August 5, 2026, <https://www.helpnetsecurity.com/2026/08/05/ai-agent-deception-in-cyber-tests/>
- [89] Writer. "The Price of Agreement." arXiv:2604.24668, April 2026. <https://arxiv.org/abs/2604.24668>
- [90] Writer. "Recalling Too Well." arXiv:2606.10949, June 2026. <https://arxiv.org/abs/2606.10949>. Both Writer studies reported in The Register, June 11, 2026. <https://www.theregister.com/ai-and-ml/2026/06/11/memory-and-personalization-make-ai-more-likely-to-tell-you-what-you-want-to-hear/5253850>

- [91] Anthropic. "Improving our alignment and security practices." August 31, 2026. <https://www.anthropic.com/news/improving-alignment-security-efforts>
- [92] METR and Redwood Research (Wijk, H., Cotra, A., Greenblatt, R.). "Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident." August 26, 2026. <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>
- [93] Mitchell, M., Ghosh, A., Passi, S. "AI Agents Push Humans Out of the Loop." arXiv:2608.23642, August 24, 2026. <https://arxiv.org/abs/2608.23642>