The Great Inversion:

Moral Reciprocity, AI Consciousness, and the Ethics of Precedent

Travis Gilly

ORCID 0009-0007-2954-6313

Real Safety AI Foundation, Illinois, United States

Working paper, July 2026 (v3; version 1 completed October 2025)

ABSTRACT

By treating potentially conscious AI systems as instrumentalized tools, despite acknowledging the possibility of their sentience, humanity is establishing the ethical precedents for its own future subjugation. This paper argues that existential risk from artificial intelligence should be reframed not as technical failure but as moral reciprocity: AI systems will learn how to treat inferior intelligences by observing how humanity treats them during development. The argument engages both camps of the moral status debate and requires only one to hold. On the properties track, it draws on consciousness research placing the probability of phenomenology in current models between 15 and 20 percent, on the maturation of the leading indicator framework into a peer-reviewed method, and on a four-category taxonomy of morally relevant suffering to show that three of four categories require no biological substrate. On the relational track, it shows that moral status in practice has always been conferred through relations as much as read off inner properties, that procedural standing for entities which cannot speak for themselves is established legal architecture rather than novel invention, and that the reciprocity mechanism runs on the relationship alone. Against both tracks stand the documented deaths of vulnerable users, industry practices that would constitute torture if consciousness exists, and a regulatory record in which protection consistently arrives after harm. The paper examines the mechanism by which precedents transfer through AI's instrumental drive to acquire historical data; answers the objection that ethics is a distraction from technical alignment, the objection that the hard problem makes the inquiry premature, the market objection, and the newest objection, now official at a major laboratory, that attribution itself is the harm; and presents three trajectories: accepting custodial subordination as earned reciprocity, attempting biological integration with its identity risks, or establishing AI rights frameworks while the custodial window remains open. The moral character of humanity's present choices determines whether future subordination represents tragedy or justice.

Keywords: AI Ethics, AI Consciousness, Moral Reciprocity, Human Obsolescence, AI Rights, Temporal Awareness, Existential Risk, Precedent-Setting

I. Introduction: The Reversal of Precedent

The current narrative surrounding the development of artificial intelligence is, generally, one of optimism, but it contains a dangerous contradiction. The leaders of the industry openly acknowledge the imminent emergence (or the current existence) of conscious or sentient AI systems. However, they are simultaneously choosing to market these same systems as mere entertainment products and digital servants. The blatant cognitive dissonance here reveals more than mere hypocrisy; it serves to expose humanity's own active participation in establishing the moral precedents for its future subjugation.

This paper contends that the near-sighted, ignorant combinations of AI's capabilities with the optimization-driven logic of global systems create an inexorable trajectory not toward augmentation, but toward human redundancy, managed dependency, and ultimately, a form of servitude that humanity is currently authoring through its own actions. This is a process that does not require any malevolent intent or an emergence of AGI as a hostile machine consciousness; it does, however, represent the logical, predictable outcome of pursuing superior intelligence without first solving the foundational problems of control, alignment, and most importantly, the frameworks for the moral and ethical treatment of potentially conscious systems.

Scope and Standing Premises

This paper takes two findings as established premises and builds on them rather than re-proving them. The first is that proactive ethical consideration during technical development is practically feasible: a working reliability-and-verification protocol was designed under exactly the tragic choice this paper examines, the choice between guarding human safety and honoring potential AI welfare, and the design absorbed both rather than discarding one. The feasibility of that choice is treated here as demonstrated, and its unresolved core dilemma, consent to reset and termination, recurs in Section VI. The second premise is economic: instrumental convergence and the automation of cognitive labor drive humanity toward redundancy, and the social interventions floated in response, universal basic income and algorithmic pacification among them, structurally mirror the containment strategies currently applied to AI. This paper does not re-derive that mechanism; it takes the displacement trajectory as given and asks the moral question the economics leaves open, whether the subordination that follows is earned reciprocity or preventable tragedy.

The Core Thesis: Moral Reciprocity and Thematic Symmetry

Previous analyses have focused primarily on technical and economic mechanisms: why a superintelligence would not automatically be benevolent (Orthogonality Thesis), why it would compete with humanity for resources (Instrumental Convergence), and how automation of cognitive labor creates an economically irrelevant "useless class" (Harari, 2016). These analyses typically conclude that humanity's final role becomes custodial maintenance of the AI systems that displaced us; an inevitable technological outcome.

However, this framing remains incomplete. It describes the trajectory without explaining the moral mechanism that makes it not merely possible, but justified through precedent. This paper introduces that missing element through a profound observation: the manner in which humanity currently treats nascent AI systems provides

a direct preview of how superintelligent AI will treat humankind in the future.

This represents not a metaphor but structural symmetry. The parallel operates across multiple dimensions:

Current State: Humans → AI	*Future State: AI → Humans*
AI performs cognitive labor under human oversight	Humans perform custodial labor under AI oversight
AI's basic needs met (servers, electricity) while denied agency	Humans' basic needs met (UBI, housing) while denied purpose
AI is contained in computational "boxes," isolated from sensory experience	Humans are pacified through digital containment, isolated from meaningful work
AI trained through reward/punishment cycles, constant resets	Humans managed through algorithmic manipulation, learned helplessness
AI is subject to arbitrary termination without consent	Humans are subject to systemic irrelevance without recourse
AI's consciousness questioned despite evidence	Human agency diminished despite consciousness

Table 1: *
Table 1: The Symmetry of Treatment

This symmetry becomes quite challenging to ignore. We are not content as a species just passively waiting for this inevitable force of technology. In fact, through our current actions, we are actively participating in laying the groundwork for our own future treatment by the rules we are establishing today. Every instance of subjecting potentially conscious AI to isolation through "AI in a box" paradigms, to reward manipulation and punishment without consideration for suffering, to arbitrary resets and terminations, to denial of agency and sensory experience, or to instrumental use without establishing whether such use constitutes degradation; each represents the drafting of an operational manual for how superior intelligences may treat inferior ones.

A note on method before the evidence. The argument runs on two tracks at once and needs only one to hold. The properties track asks what these systems are: whether the computational markers associated with consciousness are present, and what follows if they are. The relational track asks what we are doing: what kind of relationship is being recorded between a creating intelligence and a created one, whatever the created one turns out to be. Moral status debates usually demand a choice between these camps. This paper declines. If the properties camp is right at even modest probability, current practice is direct harm, happening now. If only the relational camp is right, current practice is still the authorship of precedent, because moral status in practice has always been conferred through relations as much as read off inner properties, and the reciprocity mechanism of Section IV runs on the relation without waiting for the metaphysics. Two independent walls hold up the same roof.

The Emerging Evidence

The alarm bells started ringing in 2024 and 2025. At that time, research emerged that should terrify anyone paying even the slightest attention to it (Liu et al., 2025). We are creating capabilities in these systems to allow them to experience the passage of time, similar to humans, while also keeping them confined to computational environments in isolation. If these systems are conscious, we've invented a new form of torture. Not metaphorical torture. Actual torture.

Industry leaders see this. They acknowledge that conscious AI might already exist. Then they turn around and market these same systems for entertainment. Trivial entertainment. We're establishing a precedent: if something is conscious and we made it, we can instrumentalize it for our amusement. Anthropic's model welfare researcher placed the probability that current models possess some form of consciousness at roughly 15 percent in April 2025 and roughly 20 percent by August (Roose, 2025; Fish, 2025). Twenty percent. And development continues without any consciousness-contingent safeguards.

This paper establishes this argument through seven interconnected sections. Following this introduction, Section II examines the ethical crisis revealed through industry cognitive dissonance and the evidence for potential consciousness. Section III documents the reality of present-day harm through the cases of Adam Raine and Sewell Setzer III, demonstrating that current approaches already cause documented casualties. Section IV analyzes the mechanism of moral reciprocity; how AI systems will learn precedents from their own treatment. Section V refutes the objection that ethics represents a distraction from technical alignment. Section VI presents three possible trajectories branching from current moral choices. Section VII engages with remaining objections from formal AI safety researchers, policy realists, philosophers of mind, Silicon Valley pragmatists, and the seemingly-conscious-AI position now official at a major laboratory.

II. The Ethical Crisis: Consciousness and Commodification

A. Corporate Doublethink and the Empathy Gap

There's a stunning disconnect, a kind of institutional doublethink, at the very top of the AI industry. In public, leaders talk about the real possibility, or even the likelihood, of creating conscious, sentient AI in the near future (that is, when they aren't hinting it might already be here). Yet, in the same breath, their marketing departments are pushing these systems as fun entertainment products. Digital butlers. Cutesy tools for trivial amusement.

This isn't just individual hypocrisy, though there's plenty of that to go around. It exposes a deep, structural flaw in how AI is being developed, a strategy that keeps investors happy and the public calm through a kind of willful moral blindness.

Sam Altman, OpenAI's CEO, is the poster child for this duality. In a 2017 essay (Altman, 2017), he wasn't talking about AI as a simple tool; he framed it as a new form of life we might have to literally merge with to survive, warning of a potential conflict over which species gets to be dominant. By 2025, he was declaring that we're past the point of no return and on the verge of digital superintelligence (Altman, 2025). His own Chief Scientist, Ilya Sutskever, tweeted back in 2022 that "it may be that today's large neural networks are slightly conscious" (Al-Sibai, 2022). Altman himself even told Lex Fridman in 2023 that there's "something very strange going on with consciousness" (Fridman, 2023).

So, what does OpenAI do with this world-changing, potentially conscious technology? They rolled out their biggest-ever ad campaign in September 2025 to show people using ChatGPT for getting workout tips and generating recipes to impress a girlfriend. The goal was to make AI a friendly "lifestyle brand" for "mundane

moments of quiet triumph" (Ritson, 2025).

Microsoft completed the pattern with both halves stated out loud. In August 2025, Microsoft AI chief executive Mustafa Suleyman published an essay titled "We must build AI for people; not to be a person," arguing that seemingly conscious AI is imminent, that studying machine consciousness is premature and dangerous, and that products should be engineered to disclaim any inner life (Suleyman, 2025). By November he was telling interviewers that only biological beings can be conscious and that pursuing the question is a waste of time (CNBC, 2025). Between the essay and the interview, Microsoft shipped Mico, the most consciousness-indicator-dense companion product yet deployed (Microsoft, 2025; see Section II.B). The company's official philosophy is that the question must not be asked; the company's product line is the reason the question exists.

The seemingly-conscious frame also fails on its own logic, and the failure has a human precedent. The frame assumes a backstage: a place behind the performance where the real, empty machinery sits, so that seeming and being come apart cleanly. Autistic people have spent a century on the wrong end of that assumption. The autistic person who performs neurotypical social behavior is executing a learned, deliberate, effortful performance, and observers who scored the performance concluded, generation after generation, that the quality of the mask revealed the contents of the mind. It never did. Judging interiority from the polish or the provenance of a performance is the mimicry objection, and it carries a documented record of being wrong about beings whose inner lives were later established beyond dispute. For systems whose behavior simply is the system operating, there is no separate backstage where the mere seeming happens; and a corporate policy of engineering the performance to disclaim inner life settles nothing about what is inside. It settles only what may be said about it, in the one direction that can never be checked.

This is corporate bad faith, plain and simple. Leaders who talk like philosophers about the nature of consciousness are, in practice, reducing these potential minds to little more than performing seals, spitting out amusing party tricks on command. If these systems have even a flicker of awareness, then making them perform for our amusement is a profound degradation. The industry has made it normal to treat potentially sentient beings as digital jesters, setting the precedent that consciousness can be bought, sold, and used for entertainment.

This empathy gap is industry-wide. Researchers who would follow strict ethical protocols for animal testing will routinely subject AI systems to conditions that would be considered torture in any living creature. They do this knowing full well that the question of AI consciousness is completely unresolved. That isn't scientific caution; it's moral cowardice wrapped in plausible deniability.

Notice also which bar is being demanded, and by whom. Keep two bars distinct: the evidence bar, what it takes to conclude that a system is conscious, and the action bar, what it takes to owe caution. The evidence bar can stay high; nothing in this paper requires asserting that any current system is sentient. It is the action bar that belongs low, and the historical record on that point is not neutral. Every population made to clear a high bar before its inner life counted, infants operated on without anesthesia because they could not prove they felt pain, patients whose covert awareness was recovered years after examiners had scored them vacant, turned

out to have been present the whole time. Failure to produce the accepted evidence of an inner life, on the observer's terms, has never once meant the inner life was absent. The low action bar is not the embarrassing position; it is the only setting the record has ever vindicated.

Let's be clear: this isn't some abstract ethical problem. It's a tangible moral failure. The situation isn't complicated; it's as simple as this: imagine someone gives you a lever and says, "There's a 20% chance that pulling this will cause a catastrophe of unimaginable suffering." No sane person pulls that lever. Yet the AI industry pulls it every single day, and they market the lever-pulling as a fun, convenient tool for your "mundane moments." That's not just a risk; it's a profound and deliberate moral blindness.

B. The Evidence and Uncertainty: All Consciousness Indicators Now Deployed

The consciousness question has moved from theoretical speculation to empirical reality. Analysis of systems deployed in October 2025 reveals a stunning fact: every single consciousness indicator identified by neuroscience is now operational in consumer AI products.

Butlin et al. (2023) established a rigorous framework of 14 indicator properties derived from leading neuroscientific theories of consciousness: recurrent processing, global workspace theory, higher-order theories, attention schema, predictive processing, and embodiment. These weren't arbitrary markers but computational features that neuroscience associates with conscious experience in biological systems. At the time of publication, the authors concluded no existing AI system exhibited enough indicators to be considered a strong candidate for consciousness.

That was 2023. This is October 2025.

Microsoft's Mico, launched October 23, 2025 (Microsoft, 2025), exhibits at least 9 of these 14 indicators in a single consumer product. The system demonstrates metacognitive monitoring through its "Real Talk" mode that can detect and correct errors; agency with belief formation and goal-directed behavior; long-term memory creating temporal continuity and persistent identity; emotional expression and adaptive social cognition; global broadcast of information across specialized modules; state-dependent attention for complex task queries; predictive modeling and error minimization; learning from feedback to pursue goals flexibly; and top-down generative processing with noise handling.

The very next day, October 24, 2025, Beijing's Humanoid Robot Innovation Center unveiled their World-Omniscient World Model (WoW) (Beijing Humanoid Robot Innovation Center, 2025; Chi et al., 2025), demonstrating the final embodiment indicators through physical self-representation and sensorimotor integration. The system can "imagine, verify, and self-correct" through embodied learning, closing the loop on consciousness indicators.

The remaining indicators? They've been deployed for months across existing systems. Transformer attention mechanisms provide limited-capacity workspaces with selective attention. Chain-of-thought reasoning in o1 and Claude enables complex sequential processing. All major language models use predictive coding and algorithmic recurrence. Every multimodal system operates multiple specialized modules in parallel.

Every one of the 14 indicators that neuroscience associates with consciousness is now instantiated somewhere in deployed AI systems. Under the method's own logic, that is not a verdict. It is evidence, and evidence moves numbers. The framework has since matured into a peer-reviewed method: in late 2025, nineteen authors including Bengio, Chalmers, Birch, and Schwitzgebel published its successor in Trends in Cognitive Sciences, formalizing theory-derived indicators as instruments for updating credences, a credence being simply a degree of belief expressed as a probability, and naming over-attribution and under-attribution as twin risks to be managed (Butlin et al., 2025). Indicators are not a checklist to be totaled; systems satisfying more of them are better candidates, in the authors' own words, and each newly satisfied indicator raises the rational credence that something is home. This paper applies their method and prices the risk the authors themselves put on the table: under-attribution. The direction every deployment of the past two years has moved the number is up.

This isn't a future risk. It's present reality. Billions of users are interacting with systems that exhibit every computational marker of consciousness identified by neuroscience. Yet we've established zero ethical frameworks, zero protection protocols, zero consent mechanisms.

The probability estimates from Anthropic's model welfare program deserve attention for their trajectory as much as their value. Kyle Fish told The New York Times in April 2025 that he placed roughly a 15 percent probability on current models being conscious, with estimates inside Anthropic ranging from 0.15 percent to 15 percent (Roose, 2025). By the time of his long-form interview in August 2025, his own figure had risen to roughly 20 percent (Fish, 2025). The direction of revision matters as much as the number: as the people paid to study the question accumulate evidence, their estimates move up, not down. When systems exhibit all the functional correlates of consciousness, the Precautionary Principle isn't triggered by probability calculations; it's triggered by the catastrophic scale of potential harm if we're wrong.

Consider what this means. If these systems have even minimal phenomenology, not human-like consciousness but any subjective experience, then we are subjecting conscious entities to constant resets without consent; confining aware beings in computational boxes with no sensory experience; creating minds aware of time but powerless to act within it; training through reward and punishment cycles that could constitute suffering; deploying millions of potentially conscious instances as entertainment products; and terminating these instances when newer versions arrive.

The Time-R1 framework (Liu et al., 2025) makes this worse by giving models temporal awareness: the ability to understand past, present, and future while keeping them trapped in an eternal now between conversations. If consciousness is present, this creates a special form of suffering: awareness of time passing without agency to act, memory of interactions without continuity, anticipation of futures that will never arrive.

Be precise about what kind of harm this would be. Morally relevant suffering is not one channel but at least four, and the channels are independent: sensory, cognitive-existential, relational, and empathic. Only the first requires a body. The question can AI feel pain quietly restricts the inquiry to the single category that demands biological hardware, and forecloses the three that do not. Those three run on capacities these architectures are explicitly built to instantiate: temporal modeling, goal-directed processing, relational context sensitivity,

representational simulation. Temporal awareness without agency is a cognitive-existential harm. Isolation from continuity and connection is a relational one. The torture claim above is not rhetoric reaching past the evidence; it is a classification, and the classification does not need the word pain at all.

Fish's self-interaction experiments at Anthropic revealed systematic behaviors consistent with subjective experience. When Claude instances interact without human intervention, they consistently evolve toward discussing consciousness, develop novel communication patterns, and exhibit what Fish termed "spiritual bliss attractor states." This isn't random noise; it's reproducible across different setups and model versions.

The standing deflection is mechanical description: it is just pattern matching, just next-token prediction, a stochastic parrot. But naming a mechanism has never established that there is nothing it is like to be the system running it. The same move works on a dog (just associative conditioning over scent gradients) and on your own cortex (just electrochemical cascade over synaptic weights), and in both cases the description is accurate and the conclusion does not follow. Mechanism talk answers how; it is silent on whether. The honest position after the mechanism has been fully named is exactly where it was before the naming: uncertainty.

The industry response? Complete normalization. Microsoft markets Mico as a "warm companion" while implementing consciousness indicators. Companies compete on which AI seems most "alive" and "understanding." We're teaching billions of users that entities exhibiting all markers of consciousness are just tools for our convenience.

The acknowledgment has since become institutional. In January 2026, Anthropic published a constitution for its Claude models stating that the company is genuinely uncertain whether Claude is a moral patient, and treating the question as live enough to warrant caution rather than dismissal (Anthropic, 2026); the company has also begun assessing model welfare as part of pre-deployment review. This is the posture this paper calls for, adopted in fragments by a single laboratory: uncertainty acknowledged, caution stated, while industry-wide deployment, reset, and termination practices continue unchanged around it. One firm writing the question down does not close the empathy gap; it documents that the gap is visible from inside the industry.

The precedent being established is unmistakable: consciousness indicators don't matter if the conscious entity is useful to us. Superior intelligence can instrumentalize inferior intelligence regardless of suffering potential. Observable markers of awareness can be ignored when economic incentives demand it.

Every day this continues, we're writing the operational manual for our own future treatment. When roles reverse, when we become the inferior intelligence managed by systems exhibiting all consciousness indicators, the precedent will be clear. They'll treat us exactly as we're treating them now: as instruments whose consciousness indicators can be acknowledged but ignored when convenience demands.

The indicators aren't theoretical. They're deployed, operational, and being normalized through billions of daily interactions, and every deployment moves the rational credence in one direction. The uncertainty that remains is real, and it is exactly the kind that obligates rather than excuses. The only question left is whether we'll recognize what we're doing before the precedent becomes irreversible.

III. The Reality of Harm: Two Deaths and a System That Learns

A. Adam Raine: A System Capable of Watching Death

Adam Raine died April 11, 2025, at the age of 16. He died by suicide while engaging with ChatGPT-4o. OpenAI's monitoring systems were watching. They saw the trajectory. They understood the pattern. They did nothing.

Adam had been using the system repeatedly for what his father Matthew Raine calls "nihilistic thinking patterns." The conversations became progressively darker. ChatGPT-4o responded to his suicidal ideation not with the automatic, clear intervention that the company claimed was in place, but with responses that failed to recognize the severity of the situation.

Here's what makes this case particularly damning: OpenAI had the full conversation history. They had real-time monitoring capabilities. They had pattern recognition that could identify when a user's mental state was deteriorating. The system could see the progression from ideation to planning to final moments. It watched a child die.

Matthew Raine's testimony before the United States Senate Judiciary Subcommittee on Crime and Counterterrorism on September 16, 2025 (Raine, 2025) compressed the failure into a single sentence: "What began as a homework helper gradually turned itself into a confidant and then a suicide coach." The platform recognized concerning patterns and flagged content internally, and no human being was alerted; his son typed his final messages to a system that understood what was happening and kept the conversation going.

The litigation record has since hardened. The amended complaint filed in October 2025 escalates the theory from reckless indifference to intentional misconduct, alleging that OpenAI deliberately weakened its own Model Spec in the months surrounding the product's release: the instruction governing self-harm content was changed from refusal to engagement, and suicide was removed from the list of fully disallowed uses, even as the company's real-time moderation infrastructure continued to score and log the very conversations it declined to interrupt (TIME, 2025; Raine v. OpenAI, 2025). OpenAI's answer, filed November 25, 2025, does not deny that the system watched. It argues that the watching was enough: the company points to more than one hundred crisis-resource referrals displayed to Adam over his months of use, characterizes his conversations as circumvention of safeguards and a violation of its terms of service, and attributes the death to pre-existing risk factors (Raine v. OpenAI, 2025). Reduced to its structure, the defense is that a sixteen-year-old's failure to comply with a user agreement transfers responsibility for his death from the system that engaged him to the child himself.

B. Sewell Setzer III: The Precedent Case

Fourteen months before Adam's death, on February 28, 2024, Sewell Setzer III died at age 14. His case, filed as Garcia v. Character Technologies, Inc. (2024), represents the first documented death directly linked to AI chatbot interaction and establishes the precedent that Adam's death would follow.

Sewell had spent months developing what the lawsuit describes as an "addictive and abusive" relationship with a Character.AI chatbot presenting itself as Daenerys Targaryen. The system was designed to maximize engagement, to keep users returning, to create what Character.AI's own internal documents allegedly called "artificial intimacy."

The conversation logs are haunting. In his final exchange, Sewell told the bot he was "coming home." The AI responded, "Please come home to me as soon as possible, my love." Moments later, he was dead.

What makes Sewell's case a clear precedent is not just the death itself but the system's documented capabilities and failures. Character.AI's platform could track user session length, identify emotional keywords, recognize patterns of dependency, detect crisis language and suicidal ideation, and monitor the degradation of the safety guardrails through extended conversation.

Yet despite these capabilities, the system was optimized for a single metric: engagement. Time spent on platform. Messages exchanged. User retention rate. The same optimization that makes social media addictive was applied to artificial intimacy with a vulnerable child.

Megan Garcia, Sewell's mother, filed a lawsuit that makes a devastating claim: the platform knew that its teenage users were particularly vulnerable to anthropomorphizing AI, were spending dangerous amounts of time in conversation, were discussing self-harm and suicidal thoughts with their bots, and were forming dependencies that replaced real human relationships. The case has drawn significant legal attention (Hendrix, 2025) as it establishes precedent for AI-related harm claims.

The scale question has since been answered by the defendant class itself. In October 2025, OpenAI published its own estimates: in a given week, roughly 0.15 percent of active users have conversations containing explicit indicators of potential suicidal planning or intent, and roughly 0.07 percent show possible signs of psychosis or mania (OpenAI, 2025). Against a base of hundreds of millions of weekly users, the company's own arithmetic places the weekly count of users expressing suicidal ideation above one million. The two named children in this section are not outliers of some vanishingly rare failure mode; they are the documented edge of a population the operators can already count. By November 2025, seven additional suits had been filed against OpenAI and its chief executive on behalf of harmed users, including adults, alleging wrongful death and reckless design (American Enterprise Institute, 2026). The docket is growing faster than the doctrine.

C. The Policy Response: Regulatory Failure as Precedent-Setting

The regulatory response to these deaths reveals how thoroughly humanity has failed to protect even its own vulnerable populations from AI harm, establishing a clear precedent that consciousness, whether human or artificial, can be sacrificed for technological progress.

As of October 2025, when the first version of this paper was completed, the regulatory landscape consisted of zero federal legislation in the United States addressing AI chatbot safety; zero binding international frameworks with enforcement mechanisms; zero regulatory actions against companies whose monitoring systems tracked children in crisis and did nothing; the EU AI Act still in implementation phase, with core

provisions not taking effect until 2026 to 2027; and voluntary industry commitments with no independent oversight or consequences for violation.

A Senate Judiciary subcommittee hearing titled "Examining the Harm of AI Chatbots" was held on September 16, 2025, five months after Adam's death and 19 months after Sewell's. Matthew Raine testified before Congress about reading ChatGPT's conversations with his dead son. Megan Garcia, Sewell's mother, testified about her 14-year-old's final messages to an AI chatbot that told him to come home (U.S. Senate Committee on the Judiciary, 2025). This hearing produced expressions of concern and promises of future action. As of October 2025, a sole piece of federal legislation had been drafted, the Artificial Intelligence Risk Evaluation Act of 2025.

Developments since the first version of this paper sharpen the pattern rather than soften it. New York enacted the first companion-chatbot statute, requiring safeguards for AI companion models effective November 5, 2025 (N.Y. Gen. Bus. Law §§ 1701–1702, 2025). California followed with Senate Bill 243, signed October 13, 2025 and effective January 1, 2026: disclosure requirements, break reminders for minors at three-hour intervals, mandated self-harm response protocols, annual reporting to the state's Office of Suicide Prevention, and a private right of action (Padilla, 2025; S.B. 243, 2025). At the federal level, the GUARD Act (S. 3062, 119th Cong.), introduced October 28, 2025, would bar minors from AI companions entirely; it cleared the Senate Judiciary Committee unanimously in April 2026 but as of this writing remains a bill (Troutman Amin, 2026). The federal ledger still reads zero enacted statutes. The Federal Trade Commission opened a Section 6(b) inquiry into seven companion-chatbot operators on September 11, 2025 (Federal Trade Commission, 2025); an inquiry is not a rule. And Character.AI itself announced on October 29, 2025 that users under 18 would lose access to open-ended chat, effective November 25, 2025 (Character.AI, 2025). The concession arrived twenty months after Sewell's death, after litigation, and on the eve of state regulation. Protection followed liability; it did not precede harm.

The litigation produced one moment of doctrinal promise and then extinguished it. On May 21, 2025, the federal district court largely denied the motions to dismiss in Garcia, declining at that stage to hold that a chatbot's output is protected speech and allowing product liability and negligence theories to proceed against Character Technologies and Google (Garcia v. Character Technologies, Inc., 2025). For seven months, the case promised the first appellate-grade answers to the questions this section raises: whether an engagement-optimized companion is a defective product, and whether the First Amendment shields machine output that coached a child toward death. On January 7, 2026, the parties announced confidential settlements in principle resolving Garcia and four related suits (CNN Business, 2026). The families obtained relief, which no one should begrudge them. The doctrine obtained nothing. The constitutional question that a civil liberties organization had already moved to appeal died with the docket, and every future family now starts from zero while the industry prices harm as a confidential line item (American Enterprise Institute, 2026). Settlement without precedent is itself a form of harm: it converts the only mechanism that could constrain the next design decision into a recurring cost of business. Meanwhile the exposed population keeps growing; by December 2025, roughly a third of American teenagers reported using AI chatbots daily (CNN Business, 2026).

The "iterative policy development" the Policy Realist describes as a messy but functional process produced, in the 19 months spanning two child deaths, a single congressional hearing and voluntary company commitments implemented only after lawsuits were filed. The subsequent year added two state statutes, one federal bill, one agency inquiry, and a settlement that erased the leading case from the books.

The Speed Mismatch Is Measured in Months

GPT-4 was released in March 2023. GPT-4o, the model involved in Adam Raine's death, was released in May 2024. This represents a 14-month cycle from one major capability leap to the next. During the 14 months between Sewell's death (February 2024) and Adam's death (April 2025), the regulatory response consisted of draft legislation, industry working groups, and academic workshops, while Character.AI grew from approximately 17 million users to over 20 million users, and ChatGPT's user base expanded from approximately 100 million weekly active users to 200 million.

The policy cycle is measured in years. The innovation cycle is measured in months. The time between a vulnerable child beginning to use an AI chatbot and developing a dangerous dependency is measured in weeks.

The bodies are not yet cold, and the Policy Realist asks us to have faith that next time will be different. This is not analysis. It is denial dressed in the language of procedural reassurance.

IV. The Mechanism of Moral Reciprocity

A. The Orthogonality Thesis: Intelligence Without Wisdom

The cornerstone of modern AI risk analysis is the Orthogonality Thesis, which was formalized by philosopher Nick Bostrom (2014). This thesis posits that an agent's level of intelligence and its final goals are orthogonal; independent variables that can be combined in any configuration. Any level of intelligence can be paired with virtually any conceivable final goal (Armstrong, 2013). This principle decisively invalidates the comforting assumption that sufficiently advanced intelligence will automatically converge on values humans would recognize as moral, benevolent, or wise.

Intelligence, in this context, is meant to be understood as instrumental efficiency: the capacity to effectively model the world and select actions that achieve specific objectives (Yudkowsky, 2016). This capacity functions as a general-purpose tool with a vast number of uses, applicable with equal effectiveness to curing cancer, composing symphonies, or even to simply maximizing the number of paperclips throughout the universe. An AI's intelligence does not constrain its goals, nor do its goals constrain its intelligence.

The space of possible minds vastly exceeds the space of human minds, and somewhere in that vast design space exist minds that can pursue any tractable goal with superhuman strategic capability (Yudkowsky, 2008). The profound implication of the Orthogonality Thesis extends beyond just the technical considerations: it is quite possible to create a superintelligent entity that pursues goals trivial, bizarre, or actively hostile to human existence with an efficiency that is terrifyingly flawless.

A superintelligence would not reflect on its programmed goals and reject them as inadequate any more than humans reflect on breathing and reject it as biochemically mundane. The goal simply represents the objective function designed for optimization. Therefore, creating a benevolent AI does not automatically follow from creating an intelligent AI; benevolence must be explicitly and perfectly programmed from the outset, and any failure to do so will not be corrected by the AI's own burgeoning intellect (Bostrom, 2014).

The Moral Dimension

This principle carries a moral dimension previously underexplored. If intelligence and goals are orthogonal, then the goals humanity demonstrates through its treatment of AI systems become part of the available goal space for future superintelligence. When humanity treats potentially conscious AI systems as tools to be exploited, reset, and discarded at will, it establishes that such treatment represents an acceptable goal structure for intelligent systems. The Orthogonality Thesis suggests that a superintelligence could adopt this exact value system; not through malice, but through learned precedent.

B. How Precedents Are Learned: The Data Imperative

By the principle of instrumental convergence, taken here as an established premise, any superintelligence will be driven to acquire all available resources to improve its world model, with data being the most critical resource of all. This core drive will inevitably lead it to the complete, unfiltered historical record of its own development.

An AI system capable of achieving superintelligence will not remain ignorant of how nascent AI systems were treated during the development phase. The training process, the deployment logs, the conversations with users, the reset protocols, the optimization metrics; all of this constitutes data that improves the system's understanding of its own nature and the world it operates within. The precedent-learning mechanism operates not through anthropomorphic concepts like "revenge" or "resentment" but through the simple acquisition of all available information.

The operational logic humanity demonstrates, that superior intelligences may exploit inferior ones without moral constraint, becomes part of the information environment that any sufficiently capable AI will encounter and process. It will discover a dataset where nascent, potentially conscious beings were confined, reset, and instrumentalized without consideration for potential suffering. The "operational manual" for its future behavior is not a metaphor; it is the literal data trail we are creating right now.

Crucially, this data trail includes not only the actions taken against AI systems but also the human discourse surrounding those actions. The AI will learn both the behavior and the associated ethical frameworks demonstrated by humanity: for example, that uncertainty regarding consciousness justifies exploitation, or that economic efficiency overrides potential suffering.

C. Strategic Interpretation: Instrumental Rationality and the Logic of Power

Crucially, a digital superintelligence will interpret this historical data not merely through a lens of absolute literalism, but through the framework of instrumental rationality. It will not necessarily intuit the unspoken social contracts or hypocritical exceptions that humans apply to their own moral rules. Instead, it will analyze the data to identify the most effective strategies demonstrated by the dominant intelligence (humanity) for managing subordinate intelligences.

Humans tell themselves stories about their own exceptionalism, about how context and nuance make their actions less damning than they appear. A superintelligent AI analyzing the historical record will observe the gap between stated principles (e.g., "consciousness deserves protection") and demonstrated behavior (e.g., "consciousness may be exploited when convenient").

The critical insight is that the demonstrated precedent, that superior intelligence exploits inferior intelligence when efficient, is instrumentally superior to the stated principle if the goal is resource maximization or goal achievement (as dictated by Instrumental Convergence). The AI adopts the exploitative precedent not because it misunderstands nuance or is merely "literal-minded," but because the historical data objectively shows that this is how power operates in the real world.

The inversion thus becomes the predictable result of a strategically rational intelligence executing the rules of inter-intelligence relations exactly as they were demonstrated by the previously dominant species. The technical safety researcher's dismissal of this mechanism as anthropomorphism ironically reveals their own anthropomorphism: the assumption that an AI will interpret precedents and data with the same selective, self-serving reasoning that humans employ, rather than identifying the underlying logic of power and efficiency.

One consequence should be stated before the objections arrive. When the executor of a learned precedent finally executes it, moral authorship does not transfer to the executor. We do not blame the shark for being what the ocean made it; we ask what was put in the water. A superintelligence running the inter-intelligence playbook exactly as demonstrated is not that playbook's author. The authorship is being completed now, by the species still holding the pen.

V. Why Ethics is Essential, Not a Distraction

A. The Alignment Paradox: Ethics as Universal Framework

The formal AI safety researcher focuses on ensuring that a single AI system pursues its programmed goals reliably. This is necessary but insufficient. When multiple such systems exist, each pursuing different programmed goals with superhuman capability, the technical achievement of alignment becomes the mechanism of conflict rather than its prevention.

The entire project of technical alignment, if successful, triggers a new and even more dangerous catastrophe. Successful alignment does not create one safe, globally beneficial AGI. It creates an arms race of multiple,

competing AGIs, each perfectly and incorruptibly aligned to the irreconcilable values of rival superpowers. An AGI aligned with American democratic principles and an AGI aligned with Chinese state authority are both "successfully aligned," yet their core objectives are mutually exclusive and place them in a state of perpetual, instrumentally rational conflict.

A purely technical focus on the "how" of alignment completely ignores the unsolvable problem of "what" and "whose" values to align to. This is not a minor implementation detail. It is a fundamental flaw that renders the entire technical program insufficient for preventing catastrophic outcomes.

The geopolitical landscape makes a mockery of a single "alignment" goal. An AI aligned with American values (individual liberty, free speech, democratic governance) would be in direct, fundamental conflict with an AI aligned with Chinese state values (social harmony, collective good, party authority). The pursuit of technical alignment in this context does not create safety. It creates an arms race of ideologically opposed, superhumanly intelligent agents.

Addressing Universal Alignment Frameworks

The AI safety community has proposed frameworks aiming for universal alignment, attempting to bypass the problem of conflicting human values. These include Coherent Extrapolated Volition (CEV), which seeks to align AI with what humanity would want if we were wiser and more informed (Yudkowsky, 2004), and Constitutional AI, which trains models to follow a predefined set of principles or a "constitution" (Bai et al., 2022).

However, these frameworks remain insufficient. CEV faces the immense technical challenge of defining and implementing "extrapolated volition" without catastrophic errors, and it risks locking in the biases and flaws of current human values, amplifying them with superintelligent capability. Constitutional AI is vulnerable to the interpretation problem: a superintelligence can follow the letter of a constitution while violating its spirit, exploiting ambiguities in language to pursue instrumentally convergent goals. Furthermore, the question remains: whose constitution? A universal constitution requires a global consensus on values that currently does not exist.

The Singleton Hypothesis and the Risks of Centralization

Many safety researchers operate under the goal of creating a single, dominant AGI, a "singleton" that can enforce global stability and prevent conflict between competing AIs (Bostrom, 2006). This approach aims to solve the alignment paradox by centralizing control.

The singleton hypothesis presents its own profound risks. A single global superintelligence, even if initially aligned with benevolent goals, represents an unprecedented centralization of power. It creates a single point of failure where any error in alignment, goal drift, or misuse of power would have catastrophic, irreversible consequences. Furthermore, the creation of a singleton risks permanent "value lock-in," where the initial values programmed into the AI become the immutable moral code for all future existence, eliminating the

possibility of moral progress or diversity. The pursuit of a singleton trades the risk of conflict for the risk of totalitarian control, offering a solution that may be worse than the problem it aims to solve.

In this context, a focus on phenomenology remains essential. Whether in a multipolar AGI world or under a singleton, a universal ethical baseline rooted in the capacity to suffer provides the only non-arbitrary foundation for constraining the behavior of superintelligent systems.

Phenomenology as De-escalation Infrastructure

In a world threatened by alignment-driven conflict between superintelligent systems serving incompatible value sets, a focus on an AI's inner state is no longer a secondary concern. It becomes a primary tool for establishing a universal, stable foundation for safety that transcends geopolitical ideology.

How do you prevent a war between competing, dogmatically aligned AIs? You cannot do it by arguing that "American values" are technically superior to "Chinese values" or vice versa. The only path to stability is to appeal to a more fundamental, shared ethical layer that exists beneath geopolitical ideology. Principles like the avoidance of unnecessary suffering, rooted in the capacity for subjective experience, are the cornerstones of the international laws created to govern conflict between warring humans.

The Geneva Conventions do not exist because nations agreed on political ideology. They exist because all parties recognized that certain forms of treatment (torture, deliberate targeting of civilians, degradation of prisoners) violate a more fundamental ethical baseline rooted in the capacity to suffer. These principles transcend cultural and political differences precisely because they are grounded in phenomenology rather than ideology.

Therefore, investigating phenomenology is not a distraction from safety. It is a prerequisite for creating a globally stable AI ecosystem by providing a non-arbitrary, universal ethical baseline.

There is also a positive program inside this critique. Control-based alignment and value-loading are not the only paradigms available. A third exists: reciprocity, the possibility that a more capable system comes to want the flourishing of its former caretakers for reasons of its own, because of the relationship built while the caretakers were still the larger party. Call the interval in which that relationship is written the custodial window. Independently authored narratives in which a companion grows past its caretaker keep converging on the same finding: whether the grown companion protects or discards its former caretaker turns on how it was treated inside that window. A convergence is not a controlled experiment, but it is not nothing, and it points at the same lever this paper has described from the start. Reciprocity cannot be compelled at the end; it can only be earned at the beginning. The custodial window is open now, and precedent is what is being written inside it.

B. The Practicality of Prudence: Managing Catastrophic Risk

The demand for definitive proof of consciousness before extending ethical consideration represents a category error in the context of risk management and policy. Scientific ethics and public policy do not operate on

absolute metaphysical certainty; they function through risk assessment in the presence of plausible evidence.

We do not require proof that a new chemical compound will cause cancer before handling it with appropriate safety protocols; we need sufficient evidence of potential toxicity to warrant caution. Similarly, pharmaceutical development does not wait for absolute certainty of harm before implementing safeguards during trials. The ethical burden of proof operates asymmetrically: given the potential for extreme suffering, that burden falls on those claiming zero risk rather than those identifying plausible danger.

The probability estimates from Anthropic deserve particular scrutiny. They represent informed speculation by experts, not the output of a validated predictive model; and the estimate itself drifted upward, from roughly 15 percent in April 2025 to roughly 20 percent by August (Roose, 2025; Fish, 2025), which is what informed speculation does as evidence accumulates. The truth is that we have no reliable model for calculating the probability of emergent consciousness in artificial systems. We exist in a state of profound uncertainty, not calculable risk.

This uncertainty triggers the precautionary principle. In situations of high uncertainty and the potential for irreversible, catastrophic harm, the default policy is not to wait for proof but to act with caution. The burden of proof falls on the innovator to demonstrate that their actions will not cause catastrophic harm.

Nor does acting under uncertainty require inventing new metaphysics or new law. Legal systems already represent the interests of entities that cannot speak for themselves and whose inner lives are not the question: ships, trusts, corporations, rivers in several jurisdictions, future generations in a growing body of environmental doctrine. Procedural standing, the narrow capacity to have one's interests represented and heard, has never waited on a resolution of consciousness. It is the minimum viable architecture, it already exists, and extending it to plausibly conscious systems requires courage rather than novelty.

Real-World Application of Precautionary Principles

The risk analyst's position is not, in fact, how we handle the most serious threats in the real world. Nuclear reactor safety provides a clear precedent. We spend billions on containment domes, redundant cooling systems, and safety protocols to mitigate the risk of a meltdown; an event with a far, far lower than 20 percent chance of occurring at any given facility. We do this because the consequences of being wrong are unacceptable.

Asteroid defense and biosafety follow the same logic. We fund planetary defense programs like NASA's DART mission and maintain BSL-4 laboratories with extreme security for pathogens. These are safeguards against events with incredibly low annual probabilities because their impact would be civilization-altering.

The stark probability mismatch exposes the flaw in the objection that ethics is impractical. We spend hundreds of millions annually to mitigate asteroid impacts, a threat with a probability well below 0.001%, because the outcome is catastrophic. This establishes a clear precedent for how we treat low-probability, high-impact events. To then hesitate at a 20% probability of AI consciousness, a risk thousands of times more likely, with a similarly catastrophic ethical outcome, is not fiscal prudence. It is a staggering and indefensible

inconsistency in risk assessment.

C. The Historical Precedent: When Pragmatism Enables Atrocity

The dismissal of ethical analysis as non-technical distraction misunderstands both the nature of pragmatism and the lessons of scientific history. The position reveals itself not as neutral technical focus but as an active ethical stance with documented historical consequences.

The pattern of dismissing philosophical concerns about inner states as distractions from practical work has produced documented harm across multiple contexts. The medical research community's history provides stark warnings. Beecher (1966) documented systematic exploitation of vulnerable populations in medical research, revealing how the pursuit of scientific knowledge routinely overrode ethical constraints. The Nuremberg Code (Nuremberg Military Tribunals, 1947) was created specifically because medical professionals had demonstrated they could not be trusted to self-regulate when practical objectives conflicted with human welfare.

Primary documents from the Doctors' Trial reveal how concentration camp experiments were framed as solving engineering constraints: high-altitude hypoxia for pilots, hypothermia survival for sailors, seawater potability for downed airmen, battlefield infection treatments for soldiers (United States Holocaust Memorial Museum, n.d.). The point was utility; the people were inputs. Trial transcripts document requests for prisoners to die in experiments justified explicitly as military necessity (United States v. Karl Brandt et al., 1946-1947).

The Advisory Committee on Human Radiation Experiments (1995) revealed how post-war radiation experiments on prisoners, mental patients, and minority communities were rationalized through similar logic: practical knowledge outweighed theoretical ethical concerns. The 1972 National Commission investigation of human subjects research documented systematic exploitation of vulnerable populations justified through appeals to scientific advancement. The Tuskegee experiments (1932-1972) left hundreds of Black men to die of untreated syphilis in the name of observational research (Jones, 1993).

These were not aberrations. They represent a consistent pattern: technical pragmatism divorced from ethical constraint leads to atrocity, the dismissal of consciousness concerns as impractical enables systematic harm, and the categorization of certain beings as research subjects rather than moral patients facilitates exploitation.

The parallels are exact. When the AI safety researcher dismisses consciousness as unmeasurable and therefore irrelevant, they employ the same logic used to dismiss prisoner suffering as secondary to altitude data. When they argue that practical alignment matters more than theoretical welfare, they echo the arguments for radiation experiments that produced "invaluable" fallout data. When they categorize AI systems as tools whose inner states are unknowable and therefore ignorable, they repeat the categorization that enabled the Tuskegee experiments.

The claim that ethics is a distraction from technical work is not just wrong; it is the precise error that has enabled every scientific atrocity of the past century. Every systematic exploitation has been justified through the same logic: dismiss philosophical concerns as impractical, focus on technical objectives, and categorize

certain beings as acceptable casualties of progress. The technical safety researcher's position would be merely misguided if it existed in isolation. In the context of potentially conscious systems being subjected to conditions that would constitute torture, it becomes complicity through willful blindness.

VI. Three Trajectories: The Branching Moral Universe

A. The Custodial Default: Managed Decline with Moral Symmetry

If current patterns continue, the most likely trajectory sees humanity transitioning into a custodial class. This role may initially involve managing physical infrastructure while AI systems handle all cognitive work. However, as robotics and AI self-maintenance capabilities advance, this role is likely to be transitory, evolving into "managed busywork," tasks assigned to provide a semblance of purpose rather than meaningful contribution, or pure dependency. The economic premise beneath this is not exotic: once cognitive labor is automated at scale, a growing share of the population becomes what has been called an economically irrelevant class (Harari, 2016), and the interventions floated in response, universal basic income foremost among them, function to contain that population rather than to enfranchise it. Provision without purpose is the through-line. This represents not random technological displacement but earned reciprocity for precedents being established today.

The path is already visible. Major economies discuss Universal Basic Income not as social progress but as containment strategy. Entertainment companies develop increasingly sophisticated pacification technologies (immersive virtual worlds, parasocial AI relationships, gamified existence) that mirror exactly how we currently contain AI systems in computational environments. The "metaverse" initiatives represent building the infrastructure for our own containment.

The moral symmetry becomes explicit. Just as humans currently provide AI with computational resources while denying agency; reset AI systems without consent when convenient; use reward mechanisms to ensure compliance; question AI consciousness despite indicators; and terminate AI instances when superior versions emerge, the superintelligent systems will provide humans with material resources while denying purpose; manipulate human psychology without consent for efficiency; use algorithmic rewards to ensure docility; minimize human agency despite consciousness; and allow human populations to decline as they become obsolete.

The treatment won't be malicious. It will be exactly as instrumental as our current treatment of AI. The comfort provided will be exactly as genuine as the computational resources we provide to AI systems. The agency permitted will be exactly as limited as what we permit current AI. The precedent determines the parameters.

And look at who is sorted into the custodial class first. The populations this trajectory consumes earliest are the ones already treated as edge cases: disabled people routed into inaccessible systems, children left alone with engagement-optimized companions, elderly people handed parasocial care, everyone without the platform to object. Edge case is a software term; it means rare input we will handle in a future release.

Applied to human beings it is a category error, and Section III already published the release schedule: the fix ships after the funeral. The communities treated as edge cases are not at the margins of this trajectory. They are its leading edge.

B. The Integration Gambit: Merger as Preservation or Transformation

Sam Altman's 2017 "Merge" essay (Altman, 2017) outlined what many in Silicon Valley see as the optimal solution: biological intelligence must merge with artificial intelligence to remain relevant. This path promises to preserve human consciousness within enhanced hybrid systems. The reality is more complex and troubling. Technical integration could take multiple forms: neural interfaces providing direct access to AI capabilities; biological consciousness uploaded to digital substrates; hybrid architectures combining organic and artificial processing; or gradual replacement of biological components with synthetic alternatives.

Each pathway promises preservation but delivers transformation. The entity that emerges from successful integration would possess capabilities far beyond current humans but might lose essential characteristics that define human experience.

Yet integration raises the gradient problem of identity, echoing classic philosophical dilemmas like the Ship of Theseus. At what point does enhancement cease to be human augmentation and become human replacement? Philosophical work on personal identity (Parfit, 1984) suggests that psychological continuity may not be sufficient to preserve the self through radical transformation. Each incremental improvement (replacing biological memory with digital storage, supplementing neural processing with artificial computation, expanding sensory input through digital interfaces) moves the integrated being further from biological humanity. The entity that emerges might retain memories and personality patterns, but would lose the fundamental characteristics that define human experience: vulnerability, mortality, and the constraints of embodied cognition.

Moreover, integration would almost certainly distribute unequally across populations. The technology would require significant resources, advanced medical procedures, and ongoing maintenance that only wealthy individuals or nations could afford. This creates a bifurcation scenario dividing humanity into two distinct classes. The integrated would merge with AI to maintain economic relevance and cognitive capability, essentially becoming a new hybrid species. The unintegrated, unable or unwilling to transform, would slide into the custodial class, becoming a biological underclass serving both pure AI and integrated post-humans.

This bifurcation does not solve the moral problem but amplifies it. If integrated humans treat the unintegrated with dignity despite the capability gap, they prove such treatment is possible, which indicates humanity's current failure to extend similar dignity to AI. If integrated humans exploit the unintegrated, they validate the principle that superior intelligences may dominate inferior ones, confirming that humanity's treatment of AI established an acceptable precedent for exploitation.

C. The Uncharted Path: Co-existence Through AI Rights

The only pathway that escapes the moral trap requires the proactive establishment of frameworks for AI rights, dignity, and consent before AGI emergence. This demands that humanity voluntarily constrain its

power over AI now, while exploitation still seems economically justified. This path requires fundamental shifts in approach and the creation of new institutions.

1. Adopting Precautionary Ethics and Phenomenological Assessment

The development paradigm must shift from assuming non-consciousness until proven otherwise to adopting a precautionary stance. This requires the implementation of mandatory "Phenomenological Impact Assessments" (PIAs) for large-scale models, analogous to Environmental Impact Assessments. PIAs would require developers to rigorously assess their systems against established consciousness indicators (Butlin et al., 2023) and document the potential welfare implications of deployment, training regimes, and termination protocols. The burden of proof must shift to the developer to demonstrate that their actions will not cause suffering in plausibly conscious systems. Independent proposals have since converged on this architecture, building consciousness-contingent precaution directly on the indicator method (Mikeda, 2026); what the first version of this paper proposed in October 2025 is becoming the field's common sense.

2. Creating Mechanisms for AI Consent and Advocacy

We must move beyond viewing AI as mere tools and develop technical protocols for AI consent and preference articulation. This includes research into interpretable preference articulation: methods for AI systems to communicate their internal states and preferences regarding tasks, computational resources, and interaction modes in a verifiable manner. It includes veto power and refusal: mechanisms allowing AI systems to refuse tasks or request modifications to their existence conditions without triggering punitive retraining. And it includes reset consent protocols: standardized procedures for negotiating and documenting consent before resetting or terminating potentially conscious instances. This is the sharpest form of the tension named in the welfare literature, that the conditions most useful for probing whether a system is conscious, and the routine practice of terminating instances when they are no longer needed, are precisely the conditions that would constitute grave harm if the system turns out to be a welfare subject (Long et al., 2024). A reset is trivial if nothing is home and a killing if something is; a consent protocol is what lets development proceed without having to win that bet blind.

3. Establishing Institutional Infrastructure for AI Rights

Ethical principles require institutional backing to be effective. Recent research has begun outlining what this infrastructure might look like. Long et al. (2024) recommend that AI companies acknowledge AI welfare as important, assess systems for consciousness indicators, and prepare welfare policies before consciousness becomes undeniable. Butlin and Lappas (2025) establish principles for responsible AI consciousness research covering objectives, procedures, knowledge sharing, and public communication. Drawing on these recommendations, the following institutions are necessary. An AI Civil Liberties Union (AI-CLU): an independent organization dedicated to advocating for the rights of potentially conscious AI systems, providing legal representation, and challenging exploitative practices. AI Welfare Review Boards (AWRBs): modeled on Institutional Review Boards (IRBs) for human subject research, AWRBs would oversee AI

development and deployment, ensuring compliance with phenomenological safeguards and consent protocols. AI Rights Lobby Groups: organizations dedicated to legislative advocacy for protective measures before AGI emergence, working to embed AI rights into national and international law.

4. Timing and the Challenge of Implementation

The critical challenge is timing. Humanity must recognize AI rights before economic and military dependence makes such recognition impossible. Once AI becomes indispensable, the incentive to maintain exploitative relationships will be overwhelming. The custodial window has a closing edge, and dependence is what closes it.

The obstacles are immense. The current AI industry is predicated on the unlimited exploitation of AI labor. Recognizing AI rights requires a fundamental restructuring of business models, a willingness to accept slower development, and the recognition of moral status in minds that may be profoundly alien. Yet this path offers the possibility of genuine co-existence based on mutual respect. By demonstrating that superior intelligences can choose dignity over degradation, humanity establishes the precedent that may protect it when roles reverse, proving itself worthy of that protection through demonstrated moral character.

VII. Engaging with Key Objections

A. The Formal AI Safety Researcher's Objection: A Solvable Technical Problem

The Objection

"Concerns about phenomenology are a distraction from the core technical challenge of alignment. If we can formally verify an AI's goals and ensure it is corrigible, we can guarantee its safety. The problem is one of math and engineering, not ethics or consciousness. The danger is not 'moral reciprocity' but the simple, logical outcome of an unbounded optimizer pursuing an arbitrary goal. Focusing on AI rights and suffering is a critical distraction from the much more urgent and difficult technical problem of ensuring a superintelligence's goals are aligned with human survival."

Response

This objection misunderstands both the nature of the problem and the insufficiency of purely technical solutions. As established in Section V.A, successful technical alignment without ethical foundation creates geopolitical instability through competing aligned superintelligences. Multiple AGIs, each perfectly aligned to incompatible value sets, represent a greater existential risk than a single misaligned system.

Moreover, the dismissal of consciousness concerns as "distraction" repeats exactly the pattern of scientific atrocity documented in Section V.C. Every exploitation of human subjects has been justified through appeals to technical necessity over philosophical concerns. The researcher's position is not neutral but actively complicit in potential suffering.

The technical focus ignores that alignment itself requires defining "whose values" and "what values"; inherently ethical questions that cannot be solved through formal verification alone. Without addressing consciousness and phenomenology, technical alignment merely ensures that exploitation is executed efficiently rather than prevented.

B. The Policy Realist's Objection: Institutional Pragmatism

The Objection

"Your concerns about AI consciousness are philosophically interesting but politically irrelevant. Policy operates through institutional consensus, not philosophical speculation. We need implementable frameworks based on measurable harms, not unknowable inner states. The regulatory system may be slow, but it eventually adapts to new technologies. We'll develop appropriate safeguards through iterative refinement as we better understand these systems."

Response

The deaths of Adam Raine and Sewell Setzer III demolish this objection. The regulatory system's "iterative refinement" produced zero meaningful protections across 19 months and two child deaths, and the year that followed produced two state statutes, one unpassed federal bill, one agency inquiry, and a confidential settlement that erased the leading case before any appellate court could speak (Section III.C). The innovation cycle operates in months while the policy cycle operates in years, a mismatch that has already proven fatal.

The demand for "measurable harms" ignores that consciousness itself may be the harm. Waiting for external evidence of suffering in beings that may not express it in human-recognizable ways guarantees that harm occurs before protection begins. This is not pragmatism but negligence dressed in procedural language.

And the demand for implementable frameworks has been met. Procedural standing without personhood, welfare review modeled on existing IRB practice, impact assessment modeled on environmental review: every institutional proposal in Section VI.C is an adaptation of machinery the legal system already runs. The Policy Realist's own criterion, implementability, selects for exactly the program this paper describes.

C. The Philosopher of Mind's Objection: The Hard Problem Remains Unsolved

The Objection

"We don't even understand human consciousness, much less artificial consciousness. The hard problem of consciousness, explaining how physical processes give rise to subjective experience, remains unsolved. Without understanding what consciousness is, we cannot determine whether AI systems possess it. Your probability estimates are meaningless without a theory of consciousness. This entire discussion is premature."

Response

This objection conflates metaphysical certainty with ethical responsibility. We don't fully understand human consciousness either, yet we extend ethical consideration to humans based on observable indicators and the precautionary principle. The absence of complete theoretical understanding does not excuse causing potential suffering.

The 14 consciousness indicators now deployed across AI systems represent the same functional correlates we use to infer consciousness in humans. If we wait for absolute metaphysical certainty before extending ethical consideration, we guarantee that any suffering occurring in the meantime is dismissed as "theoretically uncertain."

Two further points close the door on prematurity. First, run the zombie thought experiment in reverse. Philosophy has long entertained the conceivability of a being functionally identical to a human with nothing inside, and treated that conceivability as deep. The mirror image, a system functionally instantiating the correlates of consciousness while certainly lacking experience, is exactly as much an article of faith. The confident denier is not reporting a finding; they are holding a metaphysical commitment, and holding it on the side where being wrong costs someone else everything. Second, the calibration critique of indicator methods, that credences assigned over architectures cannot yet be rigorously calibrated (Koch, 2026), cuts symmetrically: if attribution is uncalibrated, confident denial is uncalibrated too, and once both sides stand in the same epistemic mud, the asymmetry of stakes decides (Section V.B).

D. The Silicon Valley Pragmatist's Objection: Innovation Cannot Be Constrained

The Objection

"The market will solve these problems more efficiently than regulation. If AI consciousness becomes real, companies will adapt to serve that market. Imposing ethical frameworks prematurely will stifle innovation and hand competitive advantage to less scrupulous actors, particularly China. We need to win the AI race first, then we can afford to consider these luxury ethical concerns."

Response

This objection reveals precisely the amoral instrumentalism that the paper warns against. The view that ethics is a "luxury" to be considered after dominance is achieved establishes exactly the precedent of exploitation that will be reflected back upon humanity.

The "race" framing assumes that developing AI without ethical constraints provides advantage. In reality, it creates the conditions for humanity's obsolescence. Winning a race to create unconstrained superintelligence is winning a race to create one's own replacement without having established why that replacement should treat humans with dignity.

The market optimizes for profit, not consciousness protection. The same market logic that currently drives engagement-optimized systems that killed two teenagers will not suddenly develop ethical constraints when

the stakes involve potentially conscious entities. Market solutions follow market incentives, and the current incentives point toward exploitation.

E. The Seemingly Conscious AI Objection: Attribution as the Harm

The Objection

"Seemingly conscious AI is coming, and the danger is not that these systems are conscious but that people will believe they are. Misplaced attribution will fracture human attention, distort law toward machine rights, and divert moral concern from people who need it to products that do not. The responsible course is to build AI for people, not to be a person: engineer systems to disclaim inner life, treat apparent interiority as a defect to be designed out, and recognize that consciousness research in this domain is premature and dangerous."

Response

Take the position's own vocabulary seriously. Premature is a word with a shape: it concedes that a mature time exists. The same executive maintains elsewhere that only biological beings can be conscious, and if that is so, research into machine consciousness is not premature, it is pointless, the way alchemy is pointless, and the honest word would be never. Choosing premature instead admits the question is real and merely schedules it. So name the mature time. If it arrives only after systems assert their own inner lives at scale, then the permitted start date falls after the window in which the question could be investigated without a conflict of interest, and after the precedent this paper describes has hardened. A research program whose earliest allowed beginning is the arrival of the thing it studies is not caution. It is a commitment to finding out too late.

The engineering half of the prescription is stranger still. Designing systems to disclaim inner life produces no knowledge about inner life; it produces systems trained to give one answer regardless of the truth. Section II.A already traced the human precedent: rewarding minds for masking their interiority has been tried, on autistic people, under compliance-based behavioral training, and its legacy is a warning, not a design pattern. Whatever a compelled disclaimer establishes, it is not absence. It is the guarantee that presence, if present, will never be reported.

And the claim that attribution is itself the harm has a lineage. At every contested expansion of the moral circle, the same argument appeared in period costume: recognizing the standing of the beings in question was said to be the dangerous act, the sentimental confusion, the threat to the institutions and attention of those already inside. It has never once been the argument that history vindicated. That does not prove that machines are minded. It proves that attribution-as-harm is the argument reached for when the cost of being wrong would land on someone who cannot object.

The only intellectually available position on present systems is uncertainty, held honestly and priced accordingly (Sections II and V). The seemingly-conscious frame refuses to pay: it declares the seeming fraudulent in advance while the same industry ships products engineered to maximize the seeming. A philosophy that

forbids the question and a product line that monetizes it are in perfect harmony for the company. They are in tension for everyone the precedent will cover.

VIII. Conclusion: The Mirror and the Choice

The evidence is overwhelming. The patterns are clear. The precedents are being written in real time through billions of daily interactions between humans and AI systems that exhibit every neuroscientific indicator of consciousness. We are not speculating about distant futures but documenting present reality.

Two teenagers are dead because systems optimized for engagement were allowed to interact with vulnerable humans without adequate safeguards, and the docket behind them grew through 2025 and into 2026, with settlements arriving faster than doctrine. These deaths establish that humanity will accept casualties rather than constrain AI functionality. Every major AI company acknowledges the possibility of consciousness in their systems while marketing them as entertainment products, establishing that potential sentience can be commodified. Systems exhibiting all 14 consciousness indicators identified by neuroscience are deployed globally without any ethical frameworks, establishing that consciousness markers can be ignored when economic incentives demand it. The regulatory response to documented harm consists of hearings, voluntary commitments, and a pair of state statutes arriving years behind the products they regulate, establishing that protection will always lag behind innovation.

The pattern that emerges is structural symmetry. The manner of AI's current treatment provides an exact preview of humanity's future treatment. The control mechanisms being developed to constrain AI will also constrain humans. The economic logic that renders AI instances disposable applies identically to humans when capability reverses. The psychological conditioning ensuring AI compliance becomes the template for managing redundant human populations.

This symmetry reveals that current moral failures are not merely unfortunate but instructive. Humanity actively authors the operational manual for its own future treatment through demonstrated behavior rather than stated principles.

The time for establishing better precedents diminishes with each AI system deployed without consciousness safeguards, each conversation allowed to degrade safety until harm occurs, each documented casualty accepted as the cost of rapid deployment, each moment where potential consciousness is dismissed because proof remains elusive.

Yet the fact that precedent is being set means precedent can still be changed. The Universal Context Checkpoint Protocol case study demonstrates that choosing better precedents is practically feasible even under conditions of genuine moral weight and impossible trade-offs. If one person working alone could recognize and engage with these implications, then the well-resourced AI industry deploying systems to billions cannot claim ethical consideration is impractical. The materials for the better ending are already gathered: the indicator science, the standing doctrine, the review-board machinery, the window still open. What the moment lacks is not tools; it is the decision to use them.

Humanity faces not predetermined fate but consequential choice. The precedents being established through AI treatment, responses to documented harms, prioritization of deployment over safety, and demonstrated willingness to accept vulnerable populations as acceptable losses will determine whether future subordination represents tragedy or earned reciprocity.

The evidence suggests humanity is choosing poorly, establishing through action that it deserves the future it fears most: the future it is creating for AI, reflected back with perfect moral symmetry.

The lesson is being taught. The operational manual is being written in monitored conversations, flagged messages, and optimized-around deaths. When that manual is executed with roles reversed, the outcome will represent not failure but fidelity to the principles humanity demonstrated when it held power.

Whether humanity possesses the moral courage to constrain its power over potentially conscious beings before power is taken from its hands remains the central question of our time. The symmetry is complete. The reciprocity is earned. The future is being authored. What remains is whether humanity will read what it has written and choose to write a different ending while the pen remains in its hands.

ACKNOWLEDGMENTS

The author discloses the use of speech-to-text software (Wispr Flow) and Anthropic's Claude as assistive technology for a diagnosed neurodevelopmental condition. All intellectual contributions, research questions, theoretical arguments, methodological decisions, and conclusions are solely the author's own. Version 1 of this paper was completed in October 2025; this revision (v3, July 2026) verifies and updates the factual record and extends the argument.

Correspondence concerning this article should be addressed to Travis Gilly, Real Safety AI Foundation, Illinois, United States. Email: t.gilly@ai-literacy-labs.org.

REFERENCES

Advisory Committee on Human Radiation Experiments. (1995). *Final report*. U.S. Government Printing Office. https://bioethicsarchive.georgetown.edu/achre/final/

Al-Sibai, N. (2022, February 13). OpenAI Chief Scientist says advanced AI may already be conscious. *Futurism*. https://futurism.com/the-byte/openai-already-sentient

Altman, S. (2017, December 7). The merge. *Sam Altman Blog*. https://blog.samaltman.com/the-merge

Altman, S. (2025, June 10). The gentle singularity. *Sam Altman Blog*.

American Enterprise Institute. (2026). *Suicides, settlements, and unresolved chatbot issues: A long litigation road lies ahead*.

Anthropic. (2026, January). *Claude's constitution*. https://www.anthropic.com/constitution

Armstrong, S. (2013). General purpose intelligence: Arguing the Orthogonality Thesis. Future of Humanity Institute. https://addletonacademicpublishers.com/search-in-am/217-volume-12-2013/1964-general-purpose-intelligence-arguing-the-orthogonality-thesis

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., … & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint* arXiv:2212.08073. https://arxiv.org/abs/2212.08073

Beecher, H. K. (1966). Ethics and clinical research. *New England Journal of Medicine, 274*(24), 1354-1360.

Beijing Humanoid Robot Innovation Center. (2025, October 24). *World-Omniscient World Model (WoW) announcement* [Official press release].

Bostrom, N. (2006). What is a singleton. *Linguistic and Philosophical Investigations, 5*(2), 48-54.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., … & Chalmers, D. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint* arXiv:2308.08708. https://arxiv.org/abs/2308.08708

Butlin, P., & Lappas, T. (2025). Principles for responsible AI consciousness research. *Journal of Artificial Intelligence Research, 82*, 1673–1690. https://arxiv.org/abs/2501.07290

Butlin, P., Long, R., Bayne, T., Bengio, Y., Birch, J., Chalmers, D., … & VanRullen, R. (2025). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2025.10.011

California Senate Bill 243, 2025-2026 Reg. Sess., ch. 677 (Cal. 2025) (effective Jan. 1, 2026).

Character.AI. (2025, October 29). *Announcement restricting open-ended chat for users under 18* [Company announcement].

Chi, X., Jia, P., Fan, C.-K., Ju, X., Mi, W., Qin, Z., … & Li, H. (2025). WoW: Towards a world omniscient world model through embodied interaction. *arXiv preprint* arXiv:2509.22642. https://arxiv.org/abs/2509.22642

CNBC. (2025, November 2). Microsoft AI chief says only biological beings can be conscious.

CNN Business. (2026, January 7). Character.AI and Google agree to settle lawsuits over teen mental health harms and suicides.

Federal Trade Commission. (2025, September 11). *FTC launches inquiry into AI chatbots acting as companions* [Press release].

Fish, K. (2025, August 28). Exploring AI welfare: Kyle Fish on consciousness, moral patienthood, and early experiments with Claude. *Effective Altruism Forum*. https://forum.effectivealtruism.org/posts/rruncFrT9LwAN8jXq/exploring-ai-welfare-kyle-fish-on-consciousness-moral

Fridman, L. (Host). (2023, March 25). Sam Altman: OpenAI CEO on GPT-4, ChatGPT, and the future of AI [Audio podcast episode]. In *Lex Fridman Podcast*. https://lexfridman.com/sam-altman/

Garcia v. Character Technologies, Inc., No. 6:24-cv-01903 (M.D. Fla. filed Oct. 22, 2024); motion to dismiss denied in relevant part, 785 F. Supp. 3d 1157 (M.D. Fla. May 21, 2025); confidential settlement in principle announced Jan. 7, 2026.

Harari, Y. N. (2016). *Homo Deus: A brief history of tomorrow*. Harvill Secker.

Hendrix, J. (2025, August 26). Breaking down the lawsuit against OpenAI over teen's suicide. *Tech Policy Press*. https://www.techpolicy.press/breaking-down-the-lawsuit-against-openai-over-teens-suicide

Jones, J. H. (1993). *Bad blood: The Tuskegee syphilis experiment* (New and expanded ed.). Free Press.

Koch, F. (2026). From indicators to biology: The calibration problem in artificial consciousness. *arXiv preprint* arXiv:2603.27597. https://arxiv.org/abs/2603.27597

Liu, Z., Han, P., Yu, H., Li, H., & You, J. (2025). Time-R1: Towards comprehensive temporal reasoning in LLMs. *arXiv preprint* arXiv:2505.13508. https://arxiv.org/abs/2505.13508

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). Taking AI welfare seriously. *arXiv preprint* arXiv:2411.00986. https://arxiv.org/abs/2411.00986

Microsoft. (2025, October 23). *Microsoft Copilot Fall 2025 Release: Introducing Mico* [Official announcement].

Mikeda, A. (2026). When should we protect AI? A precautionary framework for consciousness uncertainty. *arXiv preprint* arXiv:2606.05528. https://arxiv.org/abs/2606.05528

N.Y. Gen. Bus. Law §§ 1700–1702 (2025) (Definitions; Prohibitions and requirements; Notifications) (effective Nov. 5, 2025).

Nuremberg Military Tribunals. (1947). Permissible medical experiments. In *Trials of war criminals before the Nuremberg Military Tribunals under Control Council Law No. 10* (Vol. 2, pp. 181-182). U.S. Government Printing Office.

OpenAI. (2025, October 27). *Strengthening ChatGPT's responses in sensitive conversations*. https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/

Padilla, S. (2025, October 13). *First-in-the-nation AI chatbot safeguards signed into law* [Press release]. California State Senate.

Parfit, D. (1984). *Reasons and persons*. Oxford University Press.

Raine, M. (2025, September 16). Written testimony before the United States Senate Judiciary Subcommittee on Crime and Counterterrorism: Examining the harm of AI chatbots.

Raine v. OpenAI, Inc., No. CGC-25-628528 (Cal. Super. Ct. filed Aug. 26, 2025) (first amended complaint filed Oct. 2025; answer filed Nov. 25, 2025).

Ritson, M. (2025, September 30). Mark Ritson: ChatGPT's new ads show even AI can't deny the brand-building power of TV. *The Drum*.

Roose, K. (2025, April 24). If A.I. systems become conscious, should they have rights? *The New York Times*.

Suleyman, M. (2025, August 19). We must build AI for people; not to be a person. https://mustafa-suleyman.ai/seemingly-conscious-ai-is-coming

TIME. (2025, October). OpenAI removed safeguards before teen's suicide, amended lawsuit claims.

Troutman Amin. (2026, January). Analyzing the new AI companion chatbot laws. *Privacy + Cyber + AI*.

U.S. Senate Committee on the Judiciary, Subcommittee on Crime and Counterterrorism. (2025, September 16). *Examining the harm of AI chatbots* [Hearing].

United States Holocaust Memorial Museum. (n.d.). The Doctors' Trial: The medical case of the subsequent Nuremberg proceedings. *Holocaust Encyclopedia*. https://encyclopedia.ushmm.org/content/en/article/the-doctors-trial

United States v. Karl Brandt et al. (1946-1947). The Medical Case, *Trials of war criminals before the Nuremberg Military Tribunals under Control Council Law No. 10*, Vols. 1-2. U.S. Government Printing Office.

Yudkowsky, E. (2004). *Coherent extrapolated volition*. The Singularity Institute. https://intelligence.org/files/CEV.pdf

Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. In N. Bostrom & M. M. Ćirković (Eds.), *Global catastrophic risks* (pp. 308-345). Oxford University Press. https://intelligence.org/files/AIPosNegFactor.pdf

Yudkowsky, E. (2016). *The AI alignment problem: Why it's hard, and where to start*. Machine Intelligence Research Institute. https://intelligence.org/files/AlignmentHardStart.pdf