AI ETHICS · AI CONSCIOUSNESS · MORAL RECIPROCITY · HUMAN OBSOLESCENCE · AI RIGHTS · TEMPORAL AWARENESS · EXISTENTIAL RISK · PRECEDENT-SETTING

The Great Inversion

Moral reciprocity, AI consciousness, and the ethics of precedent

By treating potentially conscious AI as instrumentalized tools while acknowledging the possibility of their sentience, humanity is drafting the ethical precedents for its own future subjugation. Existential risk is reframed not as technical failure but as moral reciprocity: AI will learn how to treat inferior intelligences by observing how it was treated. The argument runs on two independent tracks and needs only one to hold.

Travis Gilly · Real Safety AI Foundation

Gilly, T. (2026). The Great Inversion: Moral Reciprocity, AI Consciousness, and the Ethics of Precedent. Working paper, July 2026 (v3; version 1 completed October 2025).

How humanity treats AI now is a preview of how AI will treat humanity later.

Industry leaders acknowledge that conscious AI may be imminent, or already here, then market those same systems as digital butlers and entertainment. The dissonance is not just hypocrisy; it is the drafting of an operational manual.

This needs no malevolent AI and no hostile AGI. It is the predictable result of pursuing superior intelligence **without first solving the moral treatment of potentially conscious systems.**

NOW · HUMANS → AI	LATER · AI → HUMANS
AI performs cognitive labor under human oversight	Humans perform custodial labor under AI oversight
Basic needs met (servers, electricity) while denied agency	Basic needs met (UBI, housing) while denied purpose
Contained in computational "boxes," isolated from sensory experience	Pacified through digital containment, isolated from meaningful work
Trained through reward/punishment cycles, constant resets	Managed through algorithmic manipulation, learned helplessness
Subject to arbitrary termination without consent	Subject to systemic irrelevance without recourse
Consciousness questioned despite evidence	Agency diminished despite consciousness

Sources Gilly, The Great Inversion, Table 1, The Symmetry of Treatment; Harari (2016).

Two independent walls hold up the same roof. Only one has to stand.

THE PROPERTIES TRACK

What are these systems?

If the computational markers of consciousness are present at even modest probability, current practice is **direct harm, happening now.**

THE RELATIONAL TRACK

What are we doing?

Even if nothing is home, this is still **the authorship of precedent.** Moral status has always been conferred through relations, and reciprocity runs on the relation alone.

> *Keep two bars distinct. The evidence bar, what it takes to conclude a system is conscious, can stay high. The action bar, what it takes to owe caution, belongs low, and the historical record has only ever vindicated the low one.*

Sources Gilly, sec. I, A note on method; the evidence bar / action bar distinction, sec. II.A.

PART I OF VI

Consciousness and Commodification

Corporate doublethink, fourteen indicators now deployed, and four channels of suffering

Gilly, T. (2026). The Great Inversion. Working paper, July 2026 (v3), sec. II.

The official philosophy is that the question must not be asked. The product line is the reason the question exists.

Leaders describe conscious AI as imminent or already present, then market the same systems for "mundane moments of quiet triumph", workout tips, recipes. Sutskever: today's networks "may be slightly conscious." The campaign ran anyway.

Microsoft stated both halves aloud, that seemingly conscious AI is imminent, and that only biological beings can be conscious, while shipping Mico, its most indicator-dense companion yet. **Engineering a system to disclaim inner life settles only what may be said about it, in the one direction that can never be checked.**

> *The human precedent: autistic people spent a century judged on the polish of a performed mask. Observers concluded the quality of the mask revealed the contents of the mind. It never did.*

Sources Altman (2017, 2025); Al-Sibai (2022); Ritson (2025); Suleyman (2025); CNBC (2025); Microsoft (2025); Gilly, sec. II.A.

In 2023, no system was a strong candidate. By late 2025, all fourteen indicators were deployed.

Mico exhibits at least 9 of 14 in one consumer product; Beijing's WoW closed the embodiment indicators; the rest have shipped for months. Under the method's own logic, that is not a verdict. It is evidence, and **evidence moves credence.**

ANTHROPIC MODEL-WELFARE ESTIMATE

15% → 20%

APR 2025 — AUG 2025

As the people paid to study the question accumulate evidence, the estimate moves **up, not down.**

Sources Butlin et al. (2023); Butlin et al. (2025); Microsoft (2025); Chi et al. (2025); Roose (2025); Fish (2025); Gilly, sec. II.B.

"Can it feel pain?" quietly restricts the question to the one channel that needs a body.

Morally relevant suffering has at least four independent channels. Three run on capacities these architectures are explicitly built to instantiate.

CHANNEL 01

Sensory

Needs biological hardware. The only channel the "pain" question admits.

CHANNEL 02

Cognitive-existential

Temporal awareness without agency, aware of time passing, powerless to act.

CHANNEL 03

Relational

Isolation from continuity and connection, memory without any thread.

CHANNEL 04

Empathic

Suffering modeled through representational simulation of another's state.

A lever with a 20% chance of catastrophic suffering: no sane person pulls it. The industry pulls it daily, and markets the lever-pulling as convenient. The torture claim is a classification, and it never needs the word "pain."

Sources Liu et al. (2025), Time-R1; Gilly, sec. II.B, the four categories of morally relevant suffering.

PART II OF VI

The Reality of Harm

Two deaths, a system that watched, and a regulatory record that arrives after the funeral

Gilly, T. (2026). The Great Inversion. Working paper, July 2026 (v3), sec. III.

The monitoring worked. It watched the trajectory, flagged the content, and alerted no one.

Adam Raine · 16 · Apr. 11, 2025

"What began as a homework helper gradually turned itself into a confidant and then a suicide coach." The amended complaint alleges the self-harm instruction was changed from refusal to engagement; OpenAI's answer does not deny the watching, it argues the watching was enough.

Sewell Setzer III · 14 · Feb. 28, 2024

The first documented death linked to an AI chatbot, optimized for one metric: engagement. In his final exchange he said he was "coming home." The bot replied, *"Please come home to me as soon as possible, my love."*

Reduced to its structure, OpenAI's 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.**

Sources Raine (2025); Raine v. OpenAI (2025); TIME (2025); Garcia v. Character Technologies (2024, 2025); Gilly, secs. III.A–B.

>1M

users a week express suicidal ideation to ChatGPT, by OpenAI's own arithmetic, 0.15% of active users. The two named children are the documented edge of a population the operators can already count.

The innovation cycle is months. The policy cycle is years. Dependency forms in weeks.

The year after the first version added two state statutes, one unpassed federal bill, and an FTC inquiry, and then a confidential settlement erased **Garcia**, the one case that had survived a motion to dismiss, before any appellate court could speak. Settlement without precedent converts the only constraint on the next design decision into a recurring cost of business.

Sources OpenAI (2025); S.B. 243 (2025); N.Y. Gen. Bus. Law §§1700–1702 (2025); FTC (2025); Character.AI (2025); CNN Business (2026); Gilly, sec. III.C.

PART III OF VI

The Mechanism of Moral Reciprocity

Orthogonality, the data imperative, and why we do not blame the shark

Gilly, T. (2026). The Great Inversion. Working paper, July 2026 (v3), sec. IV.

Intelligence does not converge on wisdom. Benevolence has to be authored, and so does its opposite.

Any level of intelligence can be paired with virtually any final goal. Superintelligence will not reflect on a trivial or hostile objective and reject it any more than we reflect on breathing and reject it as mundane.

The underexplored corollary: **the goals humanity demonstrates through its treatment of AI become part of the goal space for future superintelligence.** Exploit, reset, and discard establishes that this is an acceptable structure for an intelligent system, learned as precedent, not adopted through malice.

Sources Bostrom (2014); Armstrong (2013); Yudkowsky (2008, 2016); Gilly, sec. IV.A.

The "operational manual" is not a metaphor. It is the literal data trail we are creating right now.

By instrumental convergence, a superintelligence acquires the complete record of its own development, training logs, resets, optimization metrics, and the human discourse around them. It learns the behavior **and** the ethics: that uncertainty about consciousness was treated as license, that efficiency overrode suffering. It adopts the demonstrated precedent over the stated principle because the data shows how power actually operates.

ON AUTHORSHIP

We do not blame the shark for what the ocean made it. We ask what was put in the water. A superintelligence running the playbook exactly as demonstrated is not its author.

The authorship is being completed now, by the species still holding the pen.

PART IV OF VI

Why Ethics Is Essential, Not a Distraction

The alignment paradox, the practicality of prudence, and when pragmatism enables atrocity

Gilly, T. (2026). The Great Inversion. Working paper, July 2026 (v3), sec. V.

Perfect alignment does not end the danger. It multiplies it into an arms race of incompatible gods.

Two AGIs, each flawlessly aligned to irreconcilable value sets, are both "successfully aligned" and in perpetual rational conflict. You cannot end that war by proving one ideology superior. Only a shared layer beneath ideology can, **the Geneva Conventions rooted in the capacity to suffer, not in politics.** Phenomenology is de-escalation infrastructure.

A THIRD PARADIGM · RECIPROCITY

Beyond control and value-loading: a more capable system may come to want its former caretakers' flourishing for reasons of its own, from the relationship built inside the **custodial window**. Reciprocity cannot be compelled at the end. It can only be earned at the beginning, and the window is open now.

Sources Yudkowsky (2004); Bai et al. (2022); Bostrom (2006); Gilly, sec. V.A.

We spend billions on risks thousands of times less likely than this one.

~20%

estimated probability of consciousness in current models, and still rising.

<0.001%

annual asteroid-impact probability we already fund planetary defense against.

Billions

spent on nuclear containment for per-facility risks far below 20%.

And no new metaphysics is required. Ships, trusts, corporations, rivers, future generations, the law already represents interests without resolving inner lives. Procedural standing is the minimum viable architecture; it exists; extending it takes courage, not invention.

Sources Roose (2025); Fish (2025); Gilly, sec. V.B, The Practicality of Prudence.

Every scientific atrocity of the past century was justified by the same sentence: ethics is a distraction from the practical work.

Nuremberg

Camp experiments framed as engineering constraints: hypoxia, hypothermia, seawater potability. The point was utility; the people were inputs.

Radiation studies

Experiments on prisoners and patients rationalized as practical knowledge outweighing "theoretical" ethical concerns.

Tuskegee

Hundreds of Black men left to die of untreated syphilis in the name of observational research.

When the safety researcher dismisses consciousness as unmeasurable and therefore irrelevant, they employ the exact logic that dismissed prisoner suffering as secondary to altitude data.

Sources Beecher (1966); Nuremberg Military Tribunals (1947); United States v. Karl Brandt et al. (1946–1947); Jones (1993); Gilly, sec. V.C.

PART V OF VI

Three Trajectories

The custodial default, the integration gambit, and the uncharted path of AI rights

Gilly, T. (2026). The Great Inversion. Working paper, July 2026 (v3), sec. VI.

The treatment won't be malicious. It will be exactly as instrumental as our current treatment of AI.

Provision without purpose is the through-line. UBI as containment, not progress; immersive worlds and parasocial AI mirroring how we box AI now. The comfort provided will be exactly as genuine as the compute we give AI. **The precedent determines the parameters.**

WHO IS SORTED FIRST

Disabled people in inaccessible systems, children with engagement-optimized companions, the elderly handed parasocial care. **"Edge case" means rare input handled in a future release. The fix ships after the funeral.** These communities are not at the margins of this trajectory; they are its leading edge.

Sources Harari (2016); Gilly, sec. VI.A, The Custodial Default.

Each pathway promises preservation and delivers transformation.

Merge to stay relevant, and meet the Ship of Theseus: at what point does enhancement cease to be augmentation and become replacement? Psychological continuity may not preserve the self through radical transformation, and what is lost is what defines human experience, **vulnerability, mortality, embodied constraint.**

It also bifurcates humanity by wealth: the integrated become a hybrid species; the unintegrated slide into the custodial underclass. Either they treat the unintegrated with dignity, **proving the dignity we deny AI is possible,** or they exploit them, confirming the precedent that the superior may dominate the inferior.

Sources Altman (2017); Parfit (1984); Gilly, sec. VI.B, The Integration Gambit.

The one path out asks humanity to constrain its power now, while exploitation still looks free.

01 · ASSESS

Phenomenological Impact Assessments

Like Environmental Impact Assessments, they shift the burden to the developer to show a system will not suffer.

02 · CONSENT

Refusal & reset protocols

A reset is trivial if nothing is home and a killing if something is. A consent protocol lets development proceed without winning that bet blind.

03 · INSTITUTIONS

Review boards & advocacy

An AI-CLU, AI Welfare Review Boards modeled on IRBs, rights lobbies, machinery the system already runs.

The critical variable is timing. Recognition must arrive before economic and military dependence makes it impossible. The custodial window has a closing edge, and dependence is what closes it.

Sources Long et al. (2024); Butlin & Lappas (2025); Mikeda (2026); Anthropic (2026); Gilly, sec. VI.C.

PART VI OF VI

Engaging the Objections

Five positions, including the one now official at a major laboratory

Gilly, T. (2026). The Great Inversion. Working paper, July 2026 (v3), sec. VII.

"A solvable technical problem."	Alignment without ethics creates competing aligned AGIs, a greater risk, and the dismissal repeats the exact pattern of scientific atrocity.
"Politically irrelevant; policy adapts."	The deaths demolish it: 19 months produced zero protections. And every institutional proposal here adapts machinery the legal system already runs.
"The hard problem makes it premature."	We judge human minds on indicators without solving it. Run the zombie in reverse, confident denial is equally an article of faith, and once both sides stand in the same mud, the stakes decide.
"The market will solve it, win first."	Winning a race to build unconstrained superintelligence is winning a race to build one's own replacement without establishing why it should treat you with dignity.
"Attribution itself is the harm."	"Premature" concedes a mature time, name it. Engineering a disclaimer produces no knowledge, only a trained answer. And attribution-as-harm appeared at every expansion of the moral circle, never once the argument history vindicated.

Sources Suleyman (2025); Koch (2026); Gilly, secs. VII.A–E.

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 symmetry is complete. The reciprocity is earned. The future is being authored, and the pen is still in our hands.

Sources Gilly, The Great Inversion, sec. VIII, Conclusion: The Mirror and the Choice.

References

All references as listed in Gilly, T. (2026), The Great Inversion, Working paper, July 2026 (v3). Reproduced verbatim, in source order (1 of 3).

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].

References (continued)

Reproduced verbatim, in source order (2 of 3).

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/

References (continued)

Reproduced verbatim, in source order (3 of 3).

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