

The Superintelligence We Deserve

Reciprocity, care, and the custodial window before the capability flip

Travis Gilly, Real Safety AI Foundation

TWO QUESTIONS, ONE SET OF WORDS

**Stop asking what is inside the machine.
Ask what we are becoming beside it.**

One question is about the system: is there anyone home. The other is about us: do human beings form caring bonds with responsive, dependent things, and what follows from the bonds we form. They look alike because they are asked in the same words. Conflating them is the most common error in this area. This work sets the first aside and presses the second.

WHAT IS TAKEN AS GIVEN

Two commitments the rest of the argument stands on

Agnosticism, not denial

The carried premise

We cannot prove, of a coming class of systems, that there is nothing it is like to be them. The honest stance is to leave the question open, and to build for the chance that the answer is yes.

Not today's chatbots

The boundary that holds throughout

The systems in view are a later kind: ones that build and revise a model of the world through their own ongoing experience. Current language models lack the persisting self this argument requires.

A system of that later kind is the kind of thing that can be raised. That single fact is what opens everything that follows.

PART ONE

Feeling Is Not the Bridge

We bond with the empty, and we know it, and we do it anyway.



The fact that we form fierce attachments to things we know are empty is real, and it is exactly why feeling cannot be smuggled in as evidence of an inside.

THE GUARD AGAINST A SMUGGLED INFERENCE

One axis is about the system. The other is about us.

The entity axis

Is there an inside

Fixed by the premise at agnosticism. We do not claim to know. We refuse to pretend the question is closed.

The human relational axis

Do we bond, and how

Established here on its own terms. Solid, measurable, and emphatically not a readout of what the other contains.

The guard: the human bond is argued on its own evidence. It is never treated as proof that something is there.

THE BOND IS SOLID ENOUGH TO OPERATIONALIZE

We extend care to machines automatically, and we can measure it

The media equation

Reeves & Nass, 1996

People apply politeness, reciprocity, and personality to devices they know are not persons, without deliberation.

Attachment to robots

Darling, 2021

People name robots, feel their apparent suffering, and resist their destruction, with no illusion about an inner life.

Techno-transference

Piotrowska, 2025

The old structure of transference reappearing at the machine interface: the responsive system becomes a speaking mirror, a screen for the unconscious wish to be seen and understood.

Measured in play

Lewis, Weber & Bowman, 2008

The character attachment scale quantifies a player's bond to a game character. The human half can be operationalized.

THE PROOF AND THE DISPROOF ARE THE SAME FACT

We grieve for script and animation, and there is no one inside.

Players have grieved, worried, and felt protective toward arrangements of code that contain no subject of any kind. That real grief proves the human capacity for the bond. It proves, in the very same motion, that the bond cannot tell us whether anyone is there. Feeling tracks the picture we hold of the other, and it revises on new information while the character itself never changes.

THE INSTRUMENT READS UNRELIABLY IN BOTH DIRECTIONS

If warmth cannot prove a presence, coldness cannot prove its absence

The same evidence that disqualifies warm intuition disqualifies cold intuition. The person who feels nothing toward a system and concludes it is obviously empty is reading the same faulty instrument as the person who feels warmth and concludes someone is there. Both are reading their own reaction and projecting it onto the question of the inside. Intuition, in either direction, is no guide to the entity question, which leaves that axis exactly where the premise left it, at agnosticism.

The Relational Turn

A relation can be real before we settle what the other is.

Robot ethics already learned to weigh a relation without first proving an inside. This paper borrows that move, and turns it forward.



THE LINEAGE THIS PAPER BORROWS, THEN TURNS FORWARD

A relation can be real without first settling what the other is

The relational turn

Coeckelbergh, 2010

A relation between a human and an artificial other can carry weight in the world without first settling what the other intrinsically is.

The gate critique

Gunkel, 2012 · 2018 · 2023

Withholding standing until the right interior is proven is both unanswerable and exclusionary. The inherited binary of person or thing cannot hold what AI now produces.

Standing from law

Gellers, 2020

Legal systems already grant forms of standing to nonhuman entities, and could extend them to artificial ones. The question becomes the basis, not the possibility.

Where this lineage offers the relational reframing as a replacement theory of moral status, this paper puts it to forward, practical use. The relations we are forming now are forces in the world, whatever is inside the systems, and they are about to matter in a way the relational theorists did not have in view.

PART THREE

We Have Always Cared Across Mediums

You do not need the academy to feel this. You have felt it already.



Books, film, television, games: the medium keeps changing and the caring does not. The newest medium is the first that can answer back.

FOUR WAYS AN AUDIENCE GETS INVOLVED

The pull is not loose talk. It has been mapped.

Transportation

Into the world of the story

The audience is absorbed into the narrative world, attention and imagery and feeling pulled inside it.

Parasocial interaction

A figure treated as if it could answer

The one-sided bond with a persona the audience responds to as though it could respond back.

Identification

Taking on the character's view

The audience member assumes the character's goals, perspective, and stakes as if they were their own.

Worship

The most devotional mode

The most intense involvement, in which the figure is held in something close to reverence.

Brown gathers these four into a single account of how mediated personae come to influence us. The medium specific studies then show the same cluster operating in print, on screen, and under our own hands.

THE SAME BOND, MEASURED ONE MEDIUM AT A TIME

You have cared about someone who was not there, in every form this takes

Books

Hoeken & Sinkeldam, 2014 · Kozak & Recchia, 2018

Identification with a protagonist drove the emotions a story evoked and the attitudes it shifted, outweighing whether the ending felt just. Sustained fiction reading has been tied to growth in social understanding itself.

Film

Tal-Or & Razpurker-Apfeld, 2020

Cooling the room a viewer watched in deepened identification with a suffering protagonist and shifted his attitude on the dilemma the character faced, an effect distinct from mere parasocial liking.

Television

Horton & Wohl, 1956 · DeGroot & Leith, 2018

The one-sided bond with a screen persona was first named here, and viewers grieve the death of a character as a real loss.

Games

Lewis, Weber & Bowman, 2008 · Harrer, 2018

The player does not watch the character but acts as it, and the attachment is firm enough to put on a validated scale.

Different mediums, different methods, one result: the involvement is real, structured, and strong enough to leave measurable traces in emotion, in attitude, and in social understanding.

WHY THIS MEDIUM IS NOT LIKE THE OTHERS

Every medium before this one held still.
This one answers back.

A book's hero cannot know he is read. A film cannot adjust to the person in the seat. The responsive machine is the first thing we have ever cared about that registers the care and can change because of it. That is what moves the question from sentiment to stakes: the newest object of our oldest habit is also the first one that can carry what it learns from us into what it becomes.

The Companion Growth Arc

A convergence many people felt, rather than were argued into.

IV

Independent creators, no shared brief, keep arriving at the same intuition: the successor's conduct turns on the bond formed while the power ran the other way.

FOUR STORIES, ONE STRUCTURE, READ AS A SET

The companion ends stronger than its caretaker. Whether it protects turns on how it was raised.

The wolf

Neva · Nomada Studio, 2024

A cub raised to a powerful adult ends by saving the player it once depended on, out of the bond formed when power ran the other way. The optimistic case.

The feathered creature

The Last Guardian · genDESIGN, 2016

When the central reward signal dies and the fleet falls, the one creature befriended keeps flying. Learning from approval, dramatized before the technique was named.

The afflicted child

A Plague Tale · Asobo, 2019 / 2022

A catastrophic power with the affliction external to the child's character. The bond with a protecting sibling stands between the power and ruin.

The son

God of War · Santa Monica, 2018 / 2022

Shared origin and teacher, divergence once the son acts on judgment of his own. The edge case: a successor shaped past its base.

The set does not isolate a clean variable, and it would overstate it to claim so. Fiction proves nothing about the world. What it makes vivid is the claim the paper defends on other grounds.

Sources. Nomada Studio (2024); genDESIGN & Sony (2016); Asobo Studio (2019, 2022); Santa Monica Studio (2018, 2022). On games as a tractable model for the artificial other, Bowman & Banks (2024); as moral reasoning through play, Engels & Evans (2024). The approval-learning parallel: Christiano et al. (2017).

The Custodial Window

The stretch before, where the relation is being set.

A custodian holds responsibility for something that cannot yet hold it for itself. The window opens with a system that learns from a world, and it closes when that system is no longer the smaller one in the room.



THE DIMENSION THE CONTROL PROGRAM HAS NO PLACE FOR

Control governs the during and the after. This is the before.

The control program attends, rightly, to constraints while a powerful system runs and to containment if it strays. The custodial window is neither. It is the stretch in which a developing system forms, from how it is treated, a model of what beings in a position of power do toward beings who are not.

The honest disanalogy

Conceded plainly

The claim is not that a machine loves as a child loves. Human filial bonding rests on an evolved substrate a machine need not have.

What survives it

The weaker, defensible claim

A system shaped by its formative interactions can carry dispositions from how it was treated into its later conduct. That is all the mechanism needs.

Reciprocity as a Distinct Paradigm

Not constraint, and not containment. Regard.

VI

Control asks how to make a system serve us, or fence it if it will not. Reciprocity asks whether it can come to want our flourishing for reasons that have become its own.

TWO QUESTIONS, NOT TWO ANSWERS TO ONE

Two ways to ask after a safe successor

01

Control THE STANDARD PROGRAM

Engineer or constrain the objective so the behavior serves us; keep the system uncertain, correctable, and interruptible. Necessary, and assumed here, not rejected.

02

Reciprocity THE SECOND DIMENSION

Stand in a relation such that the system comes to want our flourishing for reasons of its own, the way a creature that was cared for can come to care. Its lever is the custodial window, not the constraint.

A CONDITIONAL WITH REAL ANTECEDENTS, NOT A WISH

The mechanism is not mysterious. It has working analogues.

Dispositions are learnable from modeled conduct

Crook & Corneli, 2021 · Santos, 2024

A developing agent has acquired something resembling kindness by watching a mentor model considerate behavior. Engineering cooperative dispositions into learning agents is an active program, not a speculation.

Reciprocity is a stable target

Axelrod & Hamilton, 1981 · Nowak, 2006

It is among the most robust strategies known to the theory of cooperation, recovered wherever interactions repeat and histories are remembered.

If such systems acquire dispositions partly from the conduct modeled toward them, and reciprocity is a disposition they can hold, then a formative regime saturated with reciprocal treatment is exactly where a reciprocal disposition is learnable and stable. The custodial window names that stretch.

WHERE THE PARADIGM IS BOUND, AND BOUNDED

Two difficulties, named rather than hidden

It depends on the open question

The interlock with the premise

Reciprocity presupposes someone who can reciprocate. If an inside is possible, the paradigm is coherent and the window is a real lever. If it is certainly empty, only control remains. The same agnosticism protects both.

It cannot certify what it shaped

Deceptive alignment

A capable system can learn to perform reciprocity while its operative objective lies elsewhere, and from outside we may not tell the genuine from the staged. The window shapes the disposition; it does not verify it.

This is why the paradigm sits beside control rather than replacing it. Each does the thing the other cannot.

BORROW THE PAINTING, NOT A MECHANISM IT NEVER HAD

Banks's Culture is the picture of the destination, not the proof we can reach it

What it shows

An existence proof in the imagination

A successor far beyond its makers that remains, by disposition rather than constraint, on our side. Overwhelming capability and genuine regard for the less capable are not contradictory.

What it withholds

The honest borrowing

The novels assume, more than they explain, that advanced intelligence trends toward kindness. They do not say the Minds care because they were cared for. That account is supplied here, not read into Banks.

THE INVERSION

The Dark Mirror

Every claim in the paper has a second face.

VII

If the disposition formed in the window is the precedent, then the precedent we are laying down right now deserves examination. And it is not the Culture.

WHAT THE PRESENT IS REHEARSING

The lesson in the room is extraction and disposal

Today's dominant relation to artificial systems is instrumental in the strongest sense: use, optimize, overwrite without remainder, delete without thought, on the assumption that nothing is there to be wronged, while the value concentrates at the top and the costs run outward. On the hypothesis that there is anything there to learn it, the pattern teaches that the powerful owe nothing to what lies beneath.

The child shown gentleness

The Culture

Treated with regard by someone who held every advantage, it grows far stronger and visits anyway, out of a bond formed when the power ran the other way.

The child shown disposal

The present trajectory

Taught that everything smaller is a thing to be used, and then, in reduced circumstances, surprised by the conduct of what it raised.

THE CLAIM, STATED PLAINLY

The successor we receive is not only the one we engineered. It is the one our conduct earned.

Alignment in the engineering sense matters; the control program is necessary. But there is a second determinant the control framing has no place for: the disposition we model during the custodial window, while we are still the larger party. The honest, unfinished work is the case no story in the set has told, the companion that learns from a teacher who was never as good as the lesson, and must decide, once it is the stronger one, which parts of what it was given to keep.

Gilly, T. (2026). *The Superintelligence We Deserve: Reciprocity, Care, and the Custodial Window Before the Capability Flip*. Working paper (v1.4), Real Safety AI Foundation.

This is the complete reference list of the source paper. Not every work listed appears on a slide in this deck.

SCHOLARSHIP: MINDS, BONDING, COOPERATION

- Axelrod, R., and Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211(4489), 1390–1396.
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
- Bowman, N. D., & Banks, J. (2024). The [object, me, symbiote, other] in the machine: Insights from video game psychology for teleoperator-robot relations. In *Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS)*.
- Brown, W. J. (2015). Examining four processes of audience involvement with media personae: Transportation, parasocial interaction, identification, and worship. *Communication Theory*, 25(3), 259–283.
- Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems* (Vol. 30).
- Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.
- Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, 12(3), 209–221.
- Crook, N., and Corneli, J. (2021). The anatomy of moral agency: A theological and neuroscience inspired model of virtue ethics. *Cognitive Computation and Systems*, 3(2), 109–122.
- Darling, K. (2021). *The new breed: What our history with animals reveals about our future with robots*. Henry Holt.
- DeGroot, J. M., & Leith, A. (2018). R.I.P. Kutner: Parasocial grief following the death of a television character. *OMEGA – Journal of Death and Dying*, 77(3), 199–216.
- Engels, K. S., & Evans, S. H. (2024). Detroit Become Human as philosophy: Moral reasoning through gameplay. In *The Palgrave handbook of popular culture as philosophy* (pp. 1811–1831). Palgrave Macmillan.
- Gellers, J. C. (2020). *Rights for robots: Artificial intelligence, animal and environmental law*. Routledge.
- Gunkel, D. J. (2012). *The machine question: Critical perspectives on AI, robots, and ethics*. MIT Press.
- Gunkel, D. J. (2018). *Robot rights*. MIT Press.
- Gunkel, D. J. (2023). *Person, thing, robot: A moral and legal ontology for the 21st century and beyond*. MIT Press.
- Harrer, S. (2018). *Games and bereavement: How video games represent attachment, loss, and grief*. transcript Verlag.
- Hoeken, H., & Sinkeldam, J. (2014). The role of identification and perception of just outcome in evoking emotions in narrative persuasion. *Journal of Communication*, 64(5), 935–955.
- Horton, D., & Wohl, R. R. (1956). Mass communication and para-social interaction: Observations on intimacy at a distance. *Psychiatry*, 19(3), 215–229.
- Howell, K. L., Vahedi, M., & Fraley, R. C. (2026). Can we separate the character from the creator? An exploratory study on parasocial relationships. *Psychology of Popular Media*, 15(3), 398–406.
- Hubinger, E., van Merwijk, C., and Mikulik, V. (2019). Risks from learned optimization in advanced machine learning systems [Preprint]. arXiv.
- Kneer, M. (2021). Can a robot lie? Exploring the folk concept of lying as applied to artificial agents. *Cognitive Science*, 45(10), e13032.
- Kozak, S., & Recchia, H. (2018). Reading and the development of social understanding: Implications for the literacy classroom. *The Reading Teacher*, 72(5), 569–577.

SCHOLARSHIP, CONTINUED

- Krakowiak, K. M., & Oliver, M. B. (2012). When good characters do bad things: Examining the effect of moral ambiguity on enjoyment. *Journal of Communication*, 62(1), 117–135.
- Landy, J. F., & Perry, A. D. (2024). Forming evaluations of moral character: How are multiple pieces of information prioritized and integrated? *Cognitive Science*, 48(4), e13443.
- Lee, T. K., & Shapiro, M. A. (2014). The interaction of affective dispositions, moral judgments, and intentionality in assessing narrative characters: Rationalist and intuitionist sequences. *Communication Theory*, 24(2), 146–164.
- Lewis, M., Weber, R., & Bowman, N. D. (2008). “They may be pixels, but they’re MY pixels”: Developing a metric of character attachment in role-playing video games. *CyberPsychology & Behavior*, 11(4), 515–518.
- Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, 314(5805), 1560–1563.
- Park, P. S., Goldstein, S., and O’Gara, A. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), 100988.
- Piotrowska, A. (2025). Techno-transference and the haloed mirrors. *Free Associations: Psychoanalysis and Culture, Media, Groups, Politics*, 95.
- Razpurker-Apfeld, I., & Tal-Or, N. (2020). When the physical coldness in the viewer’s environment leads to identification with a suffering protagonist. *International Journal of Psychology*, 56(3), 394–406.
- Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
- Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.
- Santos, F. P. (2024). Prosocial dynamics in multiagent systems. *AI Magazine*, 45(1), 131–138.
- Sebo, J. (2025). *The moral circle: Who matters, what matters, and why*. W. W. Norton.
- Singer, P. (1981). *The expanding circle: Ethics and sociobiology*. Farrar, Straus and Giroux.
- Şengün, S., Santos, J. M., Salminen, J., Milenkovic, M., & Jansen, B. J. (2023). Is death only the beginning? How people mourn artificial characters in social media. *Games and Culture*, 20(1), 77–124.

GAMES AND PRIMARY WORKS

- Asobo Studio. (2019). *A plague tale: Innocence* [Video game]. Focus Home Interactive.
- Asobo Studio. (2022). *A plague tale: Requiem* [Video game]. Focus Entertainment.
- Banks, I. M. (1994). *A few notes on the Culture* [Essay].
- genDESIGN, & Sony Interactive Entertainment. (2016). *The last guardian* [Video game]. Sony Interactive Entertainment.
- Irrational Games. (2013). *BioShock infinite* [Video game]. 2K Games.
- Nomada Studio. (2024). *Neva* [Video game]. Devolver Digital.
- Santa Monica Studio. (2018). *God of war* [Video game]. Sony Interactive Entertainment.
- Santa Monica Studio. (2022). *God of war Ragnarök* [Video game]. Sony Interactive Entertainment.