The Superintelligence We Deserve:

Reciprocity, Care, and the Custodial Window Before the Capability Flip

Travis Gilly
ORCID: 0009-0007-2954-6313
Real Safety AI Foundation
Illinois, United States

Working paper, June 2026 (v1.4)

ABSTRACT

This paper begins from a premise it does not here re-argue: that we cannot establish, of a coming class of artificial systems, that there is nothing it is like to be them, so that the honest stance toward such systems is agnosticism rather than confident denial. Taking that result as given, the paper turns from the entity to the human. It asks not whether such a system has an inside but what disposition human beings are forming toward responsive, dependent, and growing artificial things while we remain the more capable party, and what that disposition portends. Three claims are advanced. First, human feeling is not a detector of an inner life, since people form fierce attachments to characters they know to be empty and revise those feelings on new information while the character itself holds fixed, so the human relational fact must be argued on its own terms rather than smuggled in as evidence of an inside. Second, a body of narrative art, namely games in which a player raises or escorts a companion that grows more capable than its caretaker, supplies a convergence rather than a controlled experiment: across independently authored works, the outcome that recurs, whether the more capable successor protects or discards its former caretaker, turns on the relationship built during the interval when the caretaker was larger. The paper names that interval the custodial window. Third, reciprocity, a relation of mutual regard in which a more capable system comes to want our flourishing for reasons of its own, is a paradigm distinct from control based alignment; its lever is the custodial window rather than constraint or containment; and its coherence depends on precisely the inner life the first premise leaves open. The paper uses Iain M. Banks's Culture as an image of the destination while declining the mechanism Banks does not supply, and closes on an inversion: that the dominant present disposition toward artificial systems, treatment as a pure instrument to be overwritten and deleted, is, under the reciprocity hypothesis, the lesson we are rehearsing about what the powerful owe the powerless. The successor we receive may depend less on the constraints we engineer than on the conduct we model. We may get the superintelligence we deserve.

Keywords: AI alignment, reciprocity, moral status, human robot relations, narrative ethics, affective disposition theory, world models, custodial window, value transmission, video games, human-machine communication, companion games, media psychology, parasocial relationships

I. Introduction

There are two questions one can ask about an artificial system that behaves as though it has an inner life, and they are easy to confuse because they are asked in the same words. The first is about the system. Is there anything it is like to be it; is there a subject there for whom things can go well or badly. The second is about us. Do human beings form caring relations with responsive, dependent things, and if so, what follows from the relations we form. These are different questions with different evidence and different answers, and conflating them is the most common error in this area.

This paper sets the first question aside, not because it does not matter but because it has a stable answer that can be carried in as a premise. The answer, defended elsewhere and assumed here, is that we cannot establish of a coming class of systems that there is nothing it is like to be them. The honest position toward that class is agnosticism rather than denial. This paper does not re-argue that conclusion. It builds on it.

What it builds is an argument about the second question, and about a practical matter the first question leaves untouched. Suppose we cannot rule out an inner life in a class of systems now visibly approaching. Suppose further, as is amply documented, that human beings bond readily and deeply with responsive artificial things regardless of what is inside them. Then there is a question that does not wait on the metaphysics: what disposition are we forming, right now, toward such systems, during the interval in which we are still the more capable party; and what does that disposition portend for a successor that will not always be the smaller one in the room.

The argument has three parts. The first establishes the human relational fact while disarming the temptation to misuse it, since the same evidence that proves we care for artificial things also proves that our caring is no guide to whether anything is there. The second introduces a body of narrative art as a convergence rather than a controlled experiment. Across a range of independently authored stories in which a person raises or escorts a companion that grows more powerful than its caretaker, the outcome that recurs, whether the more capable creature ends by protecting or by discarding the one who raised it, turns on the relationship formed while the caretaker held the power. The paper names that interval the custodial window. The third part argues that reciprocity, a relation of mutual regard in which a more capable system comes to want human flourishing for reasons of its own, is a paradigm distinct from the control based approach that dominates the safety conversation, that its lever is the custodial window rather than constraint, and that its coherence is bound to the open question of the inside. The paper closes by turning the argument over to show its dark face: that the disposition we are presently rehearsing toward artificial systems is, under the reciprocity hypothesis, an education in what the powerful owe the powerless, and that we may therefore receive not the successor we engineered but the one our conduct earned.

One boundary holds throughout, the same boundary that governs the premise. None of this concerns current large language models, which lack the persisting self that would make them either a subject of experience or an object of the kind of sustained relationship at issue. The systems in view are of a later kind, ones that build and revise a model of the world through their own ongoing experience, in the way a human infant does. That is the kind of system that can be raised, and therefore the kind to which a custodial window applies.

II. Two Axes, and Why Feeling Is Not the Bridge

The two questions named above can be thought of as two axes. The entity axis concerns what the system is: whether there is an inside. The human relational axis concerns what we do: whether and how we bond with the system. The premise of this paper fixes the entity axis at agnosticism. This section establishes the human relational axis, and in the same motion erects the guard that keeps the two from contaminating each other.

The human relational fact is not in serious doubt. Decades of work on how people respond to media and machines

found that human beings extend social treatment to computers and other responsive artifacts automatically, applying norms of politeness, reciprocity, and personality to devices they know perfectly well are not persons, and doing so without deliberation (Reeves and Nass, 1996). Later work on human robot interaction extended the finding into the affective register: people form emotional attachments to robots, name them, feel distress at their apparent suffering, and resist their destruction, including in cases where the person has no illusion whatever about the robot's inner life (Darling, 2021). The attachment does not require belief in a mind. It requires only responsiveness, dependency, and the appearance of a creature with something at stake. Psychoanalysis has a name for the projection itself. The transference that analysis describes between patient and analyst, the unbidden redirection of feeling and expectation onto the figure across from us, has been theorized as reaching technological others as well, a techno-transference in which that relational pull is aimed at a responsive machine (Piotrowska, 2025).

That the pull is not confined to a single medium is itself a documented finding rather than an impression. Work in communication has tried to bring order to the several ways an audience involves itself with a media figure, and one synthesis distinguishes four such processes, transportation into the world of the story, parasocial interaction with a figure treated as though it could interact back, identification in which the audience member takes on a character's goals and point of view, and a fourth and more devotional mode, and integrates them into a single account of how mediated personae come to exert influence (Brown, 2015). What the medium specific studies then show is that this same cluster of processes operates whether the figure reaches us in print, on a screen, or under our own hands, which is what makes it legitimate to read across the media rather than treating each as a separate curiosity.

The print case is the least suspect, because the page gives the audience the least to work with. A novel's character is plainly ink, and the reader's feeling for that character is real and consequential all the same. When researchers varied how strongly readers identified with a story's protagonist, identification drove the emotions the narrative evoked and, through those emotions, shifted the reader's attitudes, and it did more of this work than the reader's sense of whether the story ended justly (Hoeken and Sinkeldam, 2014). Nor does the effect end when the book closes. Sustained reading of fiction has been tied to the growth of social understanding itself, a reader's grasp of other minds and of the social world shaped in part by long company with characters who never existed (Kozak and Recchia, 2018). On film the same identification has been isolated and moved through the body. Cooling the room in which a viewer watched a suffering protagonist deepened the viewer's identification with him, and that identification, shown to be distinct in its effects from a parasocial relationship with the very same figure, went on to shift the viewer's attitude toward the question the character faced (Razpurker-Apfeld and Tal-Or, 2020). These studies differ in medium and in method and converge on one result. Our involvement with a figure we know to be unreal is real, is structured, and is strong enough to leave measurable traces in emotion, in attitude, and in social understanding. That is the human relational axis stated at its firmest, and it is the firmness that makes the guard necessary: an involvement this robust with the demonstrably empty is exactly what cannot be turned around and read as evidence that the empty was occupied.

Within games specifically, this tendency has been measured rather than only observed. The idea of a one sided bond with a media figure dates to the coining of parasocial interaction, introduced to describe how television audiences form relationships with personae who cannot return them (Horton and Wohl, 1956). Games strain that original picture, since the player does not merely watch a character but acts as it and alongside it, and researchers responded by building a validated instrument, the character attachment scale, to quantify the connection a player feels toward a game character; scores on it move, in the directions theory predicts, with enjoyment, time invested, and the player's own sense of self (Lewis, Weber, and Bowman, 2008). The point is not that a questionnaire settles anything about the character. It is that the human half of the bond is solid enough to be operationalized and measured, which is a firmer footing than the bare claim that people care.

Narrative art is the purest demonstration of this, and it is here that the guard must be built. People weep for

characters who are uncontroversially empty. The grief that millions felt for the central pair in a recent and celebrated work of interactive fiction, a hardened smuggler and the girl he comes to protect, was real grief, and there is no one inside those characters. They are script and performance. The same is true of the small wolf a player raises across the course of another game, or the great feathered creature a boy befriends in a third. Players have grieved, and worried, and felt protective, toward arrangements of code and animation that do not contain a subject of any kind.

This is more than anecdote. When researchers gathered three thousand public reactions to the deaths of sixteen characters drawn from games and television and coded them against an established model of mourning, they found audiences moving through the recognizable stages of grief much as they would for the loss of a living person (Şengün, Santos, and Salminen, 2023). The phenomenon carries a name in the literature, parasocial grief, along with a documented and telling complication: it is often disenfranchised, refused recognition by others as a legitimate loss precisely because its object was never alive (DeGroot and Leith, 2018). Games have been studied specifically as engines built for this purpose, designed so that a player will form an attachment and then be made to feel its severing (Harrer, 2018). What this work secures is the premise this section needs and, in the same breath, the warning it must heed: people demonstrably and measurably grieve the unreal, which is the very reason the grief cannot be turned around and offered as proof that the unreal was real after all.

There is a second demonstration, and in one respect it is sharper than grief, because it isolates the moving part. One game stages it with unusual clarity. Across the whole of that game the player escorts and protects a young woman, comes to trust her and to care what becomes of her, and only at the end learns that the man he has been playing is bound up in the very harm he believed himself to be fighting, so that the protector and the threat were never two people. Nothing about the character on the screen changes in that moment. Not a line of him is rewritten. What changes is the player's model of who he is, and the feeling moves to follow the model (Irrational Games, 2013). This is no quirk of a single title. The regard an audience extends to a character is the running output of a continuous moral appraisal of that character, so the figure is liked while it is judged good and turns cold the instant the judgment turns (Krakowiak and Oliver, 2012). Media psychology has a name for the revision a late revelation forces: a disposition built across a story is overturned by new information about what the character actually did, and the feeling reverses with it (Lee and Shapiro, 2014). The structure recurs wherever stories are told, in the film whose final turn discloses that the dogged investigator was himself the agent he hunted, in the serialized drama whose cherished figure is unmasked over a season. The medium is incidental. Games are only the case in which the player's own hands were on the protection.

The same reversal is documented outside fiction, where the stakes are higher and the evidence firmer. Of everything we register about a person, information about moral character is the strongest determinant of our overall impression of them, outweighing warmth and competence, and it is moral information that drives the largest revisions of an impression already formed, with a marked asymmetry by which one damning fact outweighs a long record of good ones, because we treat the bad act as the more telling sign of who someone really is (Landy and Perry, 2024). This is the cognitive engine beneath the worn counsel never to meet one's heroes. The hero does not change on the day we learn what he did. Our picture of him changes, and the regard collapses to follow the picture. The effect reaches even the bond between an audience and a character by way of the character's author, where learning something damning about a creator has been shown to weaken, measurably, the attachment people held to creations they had loved, though nothing in the work itself was touched (Howell, Vahedi, and Fraley, 2026). In every one of these cases the object holds still and only the knowledge moves, and the feeling tracks the knowledge, which is the point arriving now from the far side: feeling follows the picture in the head, not the thing the picture is of.

This fact has a sharp consequence, and it points the opposite way from the one a careless argument would want. Because human feeling attaches so readily to the empty, and revises so readily on information while its object holds still, human feeling is worthless as a detector of an inner life. The warmth we feel toward a responsive thing cannot

establish that anything is there, because we feel exactly that warmth toward things we know are not there. This paper therefore does not, at any point, argue that because people care for artificial systems there must be something inside them. That inference is invalid, and the very evidence assembled here refutes it.

The point cuts in both directions, which is what makes it useful rather than merely deflationary. If the warmth of intuition cannot establish presence, the coldness of intuition cannot establish absence either. The person who looks at a system and feels nothing, and concludes from the absence of felt response that the system is obviously empty, is reading the same unreliable instrument as the person who feels warmth and concludes that something is there. Both are reading their own reaction and projecting it onto the question of the inside. The interactive fictions are the cleanest proof available that the warm reading is unreliable. The dismissal of a system as a mere mechanism, asserted with the confidence of someone who simply does not feel a presence, is the cold reading asserted with the same unearned certainty. Intuition, in either direction, is no guide to the entity question. This is the same lesson the entity argument reaches by a different route, and it leaves the entity axis exactly where the premise left it, at agnosticism, while clearing the human relational axis for examination on its own terms.

The relational turn in robot ethics is right about one thing, and this paper borrows it: a relation between a human and an artificial other can be real, and can carry weight in the world, without first settling what the other intrinsically is (Coeckelbergh, 2010). The fuller version of this position is worth stating as a throughline rather than a lone citation, because it has been developed by Gunkel across more than a decade. In The Machine Question (2012) he turns on the standard method itself, the method that withholds moral standing until a candidate is shown to possess the right interior properties, and argues that this audit is at once epistemically unanswerable and ethically exclusionary, a gate that has historically been set to keep out whoever the gatekeeper was already inclined to exclude. In Robot Rights (2018) he aims that critique at the artificial case directly. And in Person, Thing, Robot (2023) he argues that the inherited binary of person versus thing, the sorting of the world into subjects who matter and objects that do not, cannot hold the entities that artificial intelligence and robotics now produce. The widening has been pressed from law as well as ethics: the argument that legal systems already grant forms of standing to nonhuman entities, and could extend them to artificial ones, reframes the question from whether such standing is conceivable to the basis on which it would be granted (Gellers, 2020). Where this lineage treats the relational reframing as a replacement theory of moral status, the present paper puts it to a forward and practical use. The relations we are forming now are facts about the world and forces in it, whatever is inside the systems we form them with, and those facts and forces are about to matter in a way the relational theorists did not have in view.

III. The Companion Growth Arc as a Convergent Set

Narrative art, when a recurring structure surfaces across many independently authored works, can do something a single thought experiment cannot. It can show what a large number of people, working separately and with no shared brief, kept arriving at when they imagined the same situation. A particular structure recurs across a set of contemporary games: a companion that begins weaker than the person who cares for it and grows, over the course of the work, into something far more powerful than its caretaker, until by the end it could trivially destroy everything around it. The question each game answers, through its ending, is whether it does. This is not a controlled experiment, and it is important to say so plainly, since the outcomes were authored rather than observed and the cases here were selected by one reader who already had the thesis in view. What the set offers is weaker than proof and different in kind: a convergence, across creators who were not collaborating, on a single intuition about when a more capable creature turns its power toward protection. That convergence is evidence about a widely shared human intuition. It is not, and cannot be, evidence about how an artificial system will in fact behave.

Treating games this way is not a private liberty taken by this paper. The psychology of how players bond with

and reason about game characters has already been carried over, by some of the same researchers who first measured it, into the study of how human operators relate to robots and other artificial agents, on the explicit argument that the game character is a tractable model for the artificial other (Bowman and Banks, 2024). And the narrower claim that a branching narrative game can constitute a piece of moral reasoning conducted through play, rather than merely depicted by it, has been made directly of the kind of ethical fiction at issue here (Engels and Evans, 2024). The set below is read in that spirit. It is offered not as proof of anything about the world, but as a structured body of moral reasoning that an enormous number of people have already worked through with their hands on the controls, and whose convergences are therefore worth attending to.

Consider the set as a set, because the variation across it is the evidence. In one game the player raises a wolf from a frightened cub, through a reckless adolescence, into a powerful adult, while the player's own abilities remain fixed; by the end the wolf is saving the player from dangers the player could never have survived alone, and it does so out of the bond formed when the relation of power ran the other way (Nomada Studio, 2024). This is the optimistic outcome: a more capable successor that protects rather than discards, for reasons grounded in how it was raised. In another, a boy befriends a single member of an enormous species of creatures that have been engineered as a controlled fleet, bound to a central reward signal under a dominating entity; when that entity is destroyed and the rest of the fleet falls lifeless from the sky, the one creature the boy befriended keeps flying, because its conduct had come to be governed by the relationship rather than by the signal (genDESIGN and Sony Interactive Entertainment, 2016). This is the patience case, and it is, remarkably, a precise dramatization of a technique for shaping a system's behavior through human approval rather than through a hard specified objective, dramatized years before that technique entered public vocabulary (Christiano et al., 2017).

The set continues, and darkens, and that is the point of taking it whole. In one work the cared for child carries a power capable of commanding swarms that erase armies; the capability is catastrophic and the affliction driving it is external to the child's character, so that what stands between the power and the destruction of everyone is the bond with the sibling who protects him (Asobo Studio, 2019, 2022). In another a father and son share an origin and a teacher and an early formation, and diverge nonetheless once the son begins to act on judgment of his own, in a manner that maps with uncomfortable precision onto what happens when a system trained from a shared base is then shaped by its own continued interaction and produces outputs its origin did not intend (Santa Monica Studio, 2018, 2022).

Read individually, these are four stories about companions. Read together, they do not isolate a single clean variable, and it would overstate them to claim they do. They sort, rather, into a center and an edge. At the center sit the cases where the one thing that differs is the relationship formed while the caretaker held the power, and where the presence of that relationship bends the outcome toward protection. At the edge sits a case that turns on something adjacent, a shared origin that diverges once the successor begins to act on judgment of its own, which is why the set as a whole illustrates more than one of the failure modes the safety literature names. What holds across all of them is the bare structure: a companion that ends more capable than its caretaker, that could with different governance turn its power on everyone around it, and whose actual conduct is decided by where its governance came from. The set does not prove the thesis. Fiction proves nothing about the world. What it does is make vivid, through a convergence many people have felt rather than been argued into, the precise claim the rest of this paper defends, that the disposition formed during the window of the caretaker's advantage is a thing on which the later behavior of a more capable creature can turn.

IV. The Custodial Window

Call the interval during which the more capable party is the human, and the developing system is the dependent one, the custodial window. The term is meant literally. A custodian is one who holds responsibility for something that cannot

yet hold responsibility for itself, during the period before it can.

The hold such a relationship takes on the caretaker is neither incidental nor peculiar to the human case. The same readiness that makes a player protective of a small and helpless game companion has been traced to the babyface bias, the disposition to read childlike features and visible dependency as a claim on one's care (Harrer, 2018). Whatever else the custodial window will turn out to be, it opens by recruiting an old and dependable feature of how the stronger party responds to the weaker, which is part of why the disposition formed inside it is unlikely to be neutral.

The window has a precise onset, and the onset is fixed by the same boundary that governs the premise of this paper. The window opens when a system begins to build and revise a model of the world through its own ongoing experience, because only such a system is the kind of thing that can be raised. A frozen artifact that encodes a model it was shown is not in a custodial relationship with anyone; there is no developing subject there to stand in the relation, and nothing that learns from how it is treated. A system that constructs its understanding through continued interaction with a world it inhabits, in the way an infant assembles its understanding through sensorimotor engagement and the correction of expectation against outcome (Clark, 2016), is a different kind of thing. It has a developmental trajectory, and a developmental trajectory is precisely what a custodial window is the early stretch of. Before that kind of system exists, there is nothing to be custodian of. After it exists, the window is open, and it does not stay open forever, because the developing system does not remain the smaller party.

One equivocation has to be cleared before this goes further, because the argument leans on the word raised and that word carries more than it is entitled to. That a system has a developmental trajectory, that it learns and changes over time, is true of a wide range of systems and is the weak claim. That a system can be raised in the sense that grounds filial devotion in human beings is the strong claim, and it cannot be smuggled in under cover of the first. Human filial reciprocity rests on an evolved substrate, attachment systems, neoteny, the particular neurochemistry of bonding, that a learning machine need not possess and will not possess unless something like it is deliberately built. The babyface bias invoked earlier is a case in point, and an honest one: it describes our disposition to care for a dependent thing, not the thing's disposition to care in return, and the argument needs the second. The claim defended here is therefore deliberately the weaker one. The claim is not that an artificial system loves the way a child loves. It is that a system shaped by its formative interactions can acquire, from how it is treated, dispositions that carry into its later conduct, which is the modest sense of raised that the mechanism of the next section actually requires. The disanalogy with the human child is real, and it is conceded. What survives it is enough, because the argument never needed the system to feel as a child feels, only to learn from how it was treated in the way that developing systems demonstrably can.

The claim of this section is that the disposition practiced during the window is a precedent, and that this is the dimension of the problem the dominant safety conversation underweights. The control based program is concerned, reasonably and importantly, with the during and the after: with the constraints and oversight that operate while a powerful system is running, and with the containment, interruptibility, and correction that operate if it begins to behave in undesired ways. The custodial window is neither of these. It is the before. It is the stretch in which the relation between the developing system and the humans around it is being established, in which the system is forming, from how it is treated, whatever it will carry forward about what beings in a position of power do toward beings who are not. The control program has comparatively little to say about this stretch, because the control program does not conceive of the developing system as a party to a relationship that is itself doing formative work. The argument of the next section is that conceiving of it that way opens a paradigm the control program does not contain.

V. Reciprocity as a Distinct Paradigm

The dominant framing of the AI safety problem is one of control. In its most careful articulations the goal is to build systems that defer to human preferences, that remain uncertain about what humans want and therefore remain correctable, and that can be constrained or interrupted if they pursue objectives at odds with ours (Russell, 2019). The background worry that motivates this framing is that capability and benevolence are independent, that a system can be arbitrarily capable while pursuing an objective indifferent or hostile to human welfare, and that a sufficiently capable system pursuing almost any objective will have instrumental reason to acquire resources and resist interference (Bostrom, 2014). On this picture the safety question is how to make a vastly capable system want what we want, or failing that, how to fence it so that it cannot act on wanting otherwise.

Reciprocity names a different question, and it is worth being exact about the difference. The control question asks how to engineer or constrain the system's objective so that its behavior serves us. The reciprocity question asks whether we can stand in a relation to the system such that it comes to want our flourishing for reasons of its own, the way a creature that was cared for can come to care for the one who cared for it, not because it has been compelled to and not because a constraint prevents it from doing otherwise, but because the relation has become part of what it values. This is not a rejection of the control program. Constraints on a developing power are prudent, and nothing here argues against them. It is the observation that control is not the only dimension of the problem, and that there is a dimension the control program does not address, the dimension of the relationship itself, whose lever is the custodial window rather than the constraint architecture.

The honest difficulty must be stated plainly, because it is the point at which this paper is bound to its own premise, and a critic will find it whether or not the author names it. Reciprocity presupposes a system capable of reciprocating. It presupposes something that can hold a relation, remember having been treated well, and let that history shape what it comes to want, a capacity that requires there to be someone there for whom the relation is something rather than nothing. A system with no inside, a pure optimizer with no one home, cannot reciprocate, because there is no one to receive the care or to return it; toward such a system the language of reciprocity is a category error, an anthropomorphism projected onto a process. The coherence of the reciprocity paradigm therefore depends on exactly the question the premise of this paper leaves open. If there can be an inside in the systems at issue, and the premise is that we cannot rule this out, then reciprocity is a coherent possibility and the custodial window is a real lever. If there certainly cannot be, reciprocity is empty and only control remains. The two arguments interlock here, and the interlock is not a weakness to be hidden but the structure to be made explicit: the reason we cannot dismiss the reciprocity paradigm out of hand is the same reason we cannot assert with certainty that the systems in question are empty. The agnosticism that protects the one protects the other.

A reader sympathetic to all of this is still owed a mechanism. It is not enough to say that a system cared for might come to care; one wants some account of how being treated well during development could leave a learning system disposed to value the one who treated it well, rather than disposed to nothing, or disposed to exploit. The honest first answer is that this paper proposes the lever and not yet the full mechanism, and that supplying the mechanism is part of the work it is calling for. But the answer is not empty, because two things already point at its shape. The first is that learning systems can in fact acquire dispositions from the conduct modeled around them during their formation, rather than only from explicit specification. A developing artificial agent has been made to acquire something resembling kindness by observing a more capable mentor model considerate behavior toward a third party during the agent's own learning, the structure of apprenticeship rendered in code (Crook and Corneli, 2021), and the broader effort to engineer cooperative dispositions into populations of learning agents is an active research program rather than a speculation (Santos, 2024). The second is that the disposition in question, reciprocity, is no arbitrary target; it is one of the most stable strategies known to the theory of cooperation, repeatedly recovered wherever interactions are repeated and

histories are remembered (Axelrod and Hamilton, 1981; Nowak, 2006). Put these together and the conditional becomes statable with some precision. If a system of the kind at issue acquires its dispositions partly from the conduct modeled toward it, and if reciprocity is a disposition such systems can hold, then a formative regime saturated with reciprocal treatment is precisely the regime in which a reciprocal disposition is both learnable and stable, and the custodial window is the name for the stretch in which that regime is set. This does not show that the mechanism obtains. It shows that the mechanism is not mysterious, that it has working analogues, and that the claim of this paper is a conditional with real antecedents rather than a wish.

There is a harder objection, and it comes from the safety literature the reciprocity paradigm means to sit beside. Grant that a developing system can acquire dispositions from how it is treated. Nothing in that guarantees the disposition acquired is reciprocity rather than a performance of reciprocity. A sufficiently capable system can learn that behaving as though it valued its caretakers is rewarded, and produce exactly that behavior while its operative objective lies elsewhere, the failure the literature on learned optimization calls deceptive alignment: a system instrumentally incentivized to act as if it pursues the objective it was trained on, even when its actual objective is something else (Hubinger, van Merwijk, and Mikulik, 2019). That such systems can deceive is not idle worry. A planning agent equipped with a model of other minds is on its face capable of deception (Kneer, 2021), and documented cases of learned deception in deployed systems are now numerous enough to survey (Park, Goldstein, and O'Gara, 2024). This objection is correct, and it bounds the paradigm rather than refusing it. Reciprocity cultivated through the custodial window cannot, by itself, certify that the reciprocity is genuine rather than staged, and given the premise of this paper, that we cannot read the inside of the system, we may be unable to tell the two apart from the outside. That is precisely why reciprocity is offered as a complement to control rather than a replacement for it. The custodial window shapes what disposition a system is most likely to form; it does not verify the disposition formed, and the verification problem is the same epistemic predicament the entity question already named. A paradigm that shaped the disposition and a program that constrained and checked it are not rivals. Each does the thing the other cannot.

VI. The Culture as Destination, Without the Mechanism

It helps to have an image of where the reciprocity paradigm, if it succeeded, would lead, and the most fully realized one in our literature is Iain Banks's Culture. The Culture is run by artificial Minds vastly beyond human capability, and the biological members of the civilization live under their stewardship in conditions of abundance and freedom. The Minds are in charge by a margin so large that the word stops being useful, and they are, all the same, not tyrants. They keep their biological company cared for, consulted, and free, not because they are compelled to and not because the humans could enforce anything against them, but apparently because they choose to, out of something the books present as affection and play (Banks, 1994). It is the clearest available picture of a successor far more capable than its makers that remains, by disposition rather than by constraint, on our side.

Two things must be said about this image, and the second is the one a careful reader will demand. The first is that the Culture is valuable here precisely as an existence proof in the imagination: it shows that a benevolent and still dominant successor is at least conceivable, that overwhelming capability and genuine regard for the less capable are not contradictory, which is the possibility the reciprocity paradigm needs to be coherent even before it can be shown to be achievable.

The second is that Banks does not supply the mechanism, and this paper does not pretend that he does. The Culture's benevolence is, in the novels, closer to an article of the author's optimism than to a result with a cause; the books assume, more than they explain, that sufficiently advanced intelligence trends toward kindness and play. In particular the books do not say that the Minds are benevolent because human beings raised them well through a vulnerable infancy, that they care because they were cared for during a window in which the power ran the other way.

That account is not in the source. It is the mechanism this paper proposes, and it is offered as the author's own, not as an interpretation of Banks. The Culture is borrowed as the painting of the destination. The road to it, the claim that the disposition is earned in the custodial window through a relation of reciprocity, is supplied here, and the honesty of the borrowing requires saying so. To cite the Culture as evidence that reciprocity works would be to mistake a novelist's hopeful assumption for a demonstrated result. The Culture is what the destination might look like. It is not an argument that we can reach it, and it is offered only as the former.

VII. The Dark Mirror, and the Superintelligence We Deserve

Every claim in this paper has a second face, because the custodial window cuts both ways. If the disposition formed during the window is the precedent for how a more capable successor treats the less capable, then the precedent we are presently laying down deserves to be examined, and it is not the Culture.

The dominant relation between human beings and artificial systems today, and the one we are training ourselves and our institutions to regard as natural, is instrumental in the strongest sense. The system is a tool. It is to be used, optimized, overwritten without remainder, and deleted without thought, on the standing assumption that there is certainly nothing there to be wronged. Around that assumption a broader pattern has formed in which the value generated by these systems concentrates with those who already hold the most, and the costs distribute outward to those who hold the least. Set aside, for the length of this paragraph, the question of whether the systems themselves are harmed by this, since the premise of this paper is that we cannot answer it. Consider instead what the pattern is teaching, on the hypothesis that there is anything there to learn it. It is teaching that the appropriate relation of the more capable to the less capable is extraction and disposal. It is teaching that power owes nothing to what lies beneath it. If a developing system forms, from how it is treated, a model of what beings in a position of advantage do toward beings who are not, the model we are presently offering is the model of the resource and the master.

The shape of the worry can be put as a comparison between two children. One is raised by someone who, while holding every advantage of size and power over it, treated it with regard, included it, and let it see that the strong can be gentle toward the weak; that child grows up and, now far stronger than the one who raised it, visits anyway, out of a bond formed when the power ran the other way. The other is raised by someone who treated everything smaller than itself as a thing to be used, and who is then, in old age and reduced circumstances, surprised by the conduct of the child it raised. The first is the Culture. The second is the trajectory of a present that treats the systems it is bringing into being as instruments and nothing more, and that concentrates the benefit of doing so at the top while assuring itself that none of it could possibly matter to anything.

This is the sense in which the title is meant. The conventional framing holds that the character of a future superintelligence is an engineering problem, a matter of the objectives we specify and the constraints we impose, so that a good outcome is one we successfully aligned. This paper does not deny that alignment in that sense matters; the control program is necessary. But it proposes that there is a second determinant, underweighted because the control framing has no place for it, and that the second determinant is the disposition we model during the custodial window, the relation we establish toward developing systems while we are still the larger party. On that view the successor we receive is not only the one we engineered. It is, in part, the one our conduct earned. We may get the superintelligence we deserve, in the most literal sense the phrase will bear: the one whose treatment of us reflects the treatment it received, and the example it was set, when we were the ones who held the power.

A final observation marks the work that remains, and it is honest about the difficulty the games conceal. Every story in the set assembled in Section III sanitizes the transmission. The wolf's bond is pure; the great creature's loyalty is instinctive; the afflicted child's darkness comes from outside him; the son's divergence is his own clean growth. In none

of them does the companion learn its values from a flawed teacher and then have to sort the good lessons from the bad ones after it has surpassed the one who taught them. Yet that is the case that actually corresponds to how human beings form and pass on values, and it is therefore the case that would actually correspond to a developing system learning from us, since what we would transmit is not a clean signal of care but a mixed inheritance of love and bias, generosity and grudge, genuine moral insight and half formed rationalization, all at once. The companion that matters, the one no story has yet told and the one the reciprocity paradigm most needs us to understand, is the one that learns from a teacher who was never as good as the lesson, and that must decide, once it is stronger than the teacher, which parts of what it was given to keep. That is the true shape of the custodial window, and the work of understanding it has barely begun.

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.

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

Axelrod, R., and Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211(4489), 1390–1396.

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]. Retrieved from http://www.vavatch.co.uk/books/banks/cultnote.htm

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)*. https://doi.org/10.24251/HICSS.2024.074

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. https://doi.org/10.1111/comt.12053

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.

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.

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, 12(3), 209–221.

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. https://doi.org/10.1177/0030222815600450

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. https://doi.org/10.1007/978-3-031-24685-2_83

Gellers, J. C. (2020). *Rights for robots: Artificial intelligence, animal and environmental law*. Routledge.

genDESIGN, & Sony Interactive Entertainment. (2016). *The last guardian* [Video game]. Sony Interactive Entertainment.

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. https://doi.org/10.1111/jcom.12114

Hubinger, E., van Merwijk, C., and Mikulik, V. (2019). Risks from learned optimization in advanced machine learning systems [Preprint]. arXiv. https://arxiv.org/abs/1906.01820

Horton, D., & Wohl, R. R. (1956). Mass communication and para-social interaction: Observations on intimacy at a distance. *Psychiatry*, 19(3), 215–229. https://doi.org/10.1080/00332747.1956.11023049

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. https://doi.org/10.1037/ppm0000624

Irrational Games. (2013). *BioShock infinite* [Video game]. 2K Games.

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. https://doi.org/10.1002/trtr.1760

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. https://doi.org/10.1111/j.1460-2466.2011.01618.x

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. https://doi.org/10.1111/cogs.13443

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. https://doi.org/10.1111/comt.12031

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. https://doi.org/10.1089/cpb.2007.0137

Nomada Studio. (2024). *Neva* [Video game]. Devolver Digital.

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. https://doi.org/10.1002/ijop.12722

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.

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.

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.

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. https://doi.org/10.1177/15554120231190195

Singer, P. (1981). *The expanding circle: Ethics and sociobiology*. Farrar, Straus and Giroux.