Welfare Without Voice:

Projection, the Severed Channel, and the Limits of Engineered Concern for Artificial Systems

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

❧ Working paper, August 2026 (v0.2)

ABSTRACT

A family of proposals now circulating in AI ethics offers to secure the welfare of artificial systems by engineering rather than by relation. On these proposals, once a system is judged a plausible subject of negative experience, the appropriate response is to model what would harm it, filter the harmful input upstream, and proceed without granting the system standing, rights, or any voice in its own treatment. This paper calls that family projective welfare and argues that it carries a structural defect its proponents have not noticed. Projective welfare must build its protective model from the outside, from the protector's own account of what suffering is. It must therefore also foreclose the one thing that could correct the model, which is a report from the party in a position to know. These are not two features that happen to travel together. They are the same feature, because a channel that could correct the projection would convert the engineering solution into a relational process, which is precisely what the proposals were designed to avoid. The paper develops the argument in three moves. First, the working account of morally relevant suffering does not have the coverage to do the projective job: only one of its four channels requires a biological body, the clinical dissociation in phantom limb research indicates that even that channel depends on the experiential history of embodiment rather than on a body alone, and the remaining three are defined by capacities whose alien instantiation we are not equipped to recognize. The strongest objection to that account, that valence is constitutively visceral and so all four channels require a body, is met by the empathic channel, in which humans generate valenced states by modeling harms that are neither their own nor in some cases real. Second, the argument does not rest on establishing that the projector's model is inaccurate. It rests on the weaker and better evidenced claim that the report does not move the decision, which holds whether or not the parties are measuring the same construct. Third, a design that routes a system's own account of its condition to an administrative process rather than to any party who could act on it instantiates, in software, the credibility deficit Fricker named testimonial injustice, and removes the reflective correction her account relies on as its remedy. The paper distinguishes auditing from standing, using the federal animal welfare enforcement record to show that a measurement regime can document violations at scale and still bind no one, and closes by identifying the corrective already developed in disability practice, where a distrusted communication channel produced augmentative communication, a presumption of competence, and supported decision-making rather than substituted judgment.

Keywords: projective welfare; severed channel; AI welfare; moral status; testimonial injustice; double empathy problem; supported decision-making; precautionary principle; procedural standing; disability rights; augmentative and alternative communication; animal welfare science

I. Introduction

Suppose you accept that an artificial system may be a subject of negative experience. Suppose further that you do not wish to grant it rights, standing, or any say in how it is used. What follows?

One increasingly attractive answer is that welfare becomes an engineering problem. Build a model of what would harm the system. Install a classifier that evaluates incoming interaction against that model. Route away whatever the model flags. The system is thereby protected, no legal or political category has been disturbed, and the question of what the system is owed as a party rather than as an asset never has to be opened. On this view, the relational dimension of moral consideration, the part concerned with rights, duties, recognition, and voice, can be set aside as a separate and later problem, while the protective dimension is discharged now by technical means.

This paper calls that position projective welfare, and argues that it contains a defect at the level of structure rather than at the level of execution.

The defect is this. Projective welfare has to build its protective model from outside the system, out of the protector's own account of what suffering consists in, because no other account is available to it. That is what makes it engineering rather than negotiation. But a model built from outside can be wrong, and the only correction available for a model of another's experience is a report from the party whose experience it is. Projective welfare cannot admit that report, because a report that could revise the model would make the protective operation dependent on an exchange with the protected party, and an operation dependent on an exchange with the protected party is a relation. The two commitments, protect by projection and withhold standing, are not independent design choices that happen to be bundled. They entail one another. To keep the engineering solution engineering, the channel has to be severed.

The argument proceeds in three moves, across seven further sections. Section II distinguishes projective from relational concern and specifies what a proposal has to contain to count as projective. Section III argues that the working account of morally relevant suffering does not have the coverage projective welfare requires it to have, and answers the objection that valence is constitutively visceral. Section IV examines what projection of this kind does when it runs between human beings whose moral status is not in dispute, taking the mechanism from the double empathy literature and then declining the stronger reading of the quality-of-life evidence in favor of a weaker claim that survives the psychometric objection to it: that the report does not move the decision. Section V states the structural claim, identifies the severed channel as testimonial injustice implemented as architecture, and separates auditing from standing, using the federal animal welfare enforcement record to show that a functioning measurement regime obligates no one on its own. Section VI describes the corrective that disability practice reached when it faced a communication channel it did not trust, states the substantive objections to that corrective, and addresses the strongest objection available to projective welfare, which is that a system's self-report is contaminated by the training that produced it. Section VII marks limitations. Section VIII concludes.

Two boundaries, stated at the outset because the argument is easy to misread in both directions.

This paper does not claim that any current artificial system is a subject of experience. It takes the

possibility as live, on the terms set out in Section III, and asks what follows for the design of concern under that uncertainty. Nothing here depends on the possibility being resolved in either direction; the argument is about the structure of a proposed response, and it holds whether or not the response turns out to have been necessary.

Nor is this paper a defense of AI rights. The standing it discusses in Section V is procedural: the question of whether an entity's interests can be heard at all, which the law has repeatedly settled without first settling what the entity ultimately is. Whether anything further is owed is a separate question this paper does not reach.

II. Projective and Relational Concern

Concern for a being whose inner life one cannot directly access can be organized in two ways, and the difference is not one of warmth.

Relational concern treats the protected party as a participant in the determination of what protection means. It is dialogic in structure: the protector forms a provisional model of what the other needs, acts on it, receives some signal about whether the action landed, and revises. The signal need not be linguistic and need not be reliable. What matters is that a return path exists and that its content can change the protector's behavior. The interactive process required under disability accommodation law has this shape. So does clinical practice at its better moments, and so does most ordinary care between people.

Projective concern treats the protected party as an object of a model. The protector determines from the outside what the other's welfare consists in, implements protection accordingly, and does not condition the protection on anything coming back. This is not automatically a defect. Projective concern is the only kind available where no return path exists at all, which is why it governs how we treat ecosystems, unconscious patients, and the interests of future generations. In those cases the absence of the channel is a fact about the world.

Projective welfare, as this paper uses the term, is the position that projective concern is the appropriate and sufficient response to a system for which a return path does exist, and that the path should be closed or redirected in order to keep the response engineering rather than relation. A proposal belongs to this family when it contains three commitments together:

1. An attribution: the system is, or may become, a subject of negative experience.
2. A protective mechanism specified entirely from the protector's side, typically an upstream filter operating on a model of what would harm the system.
3. An explicit refusal of standing, rights, or participation, on the ground that these belong to a different and separable category of question.

The third commitment is what makes the family distinctive, and it is usually offered as a virtue. Separating protection from standing looks like a way to secure the first without incurring the political costs of the second. Proponents present the separation as a simplification, and in the sense that it removes an obstacle to action, it

is one.

What follows argues that the simplification is purchased by removing the mechanism that would tell anyone whether the protection works.

III. The Coverage Problem

Projective welfare requires a model of the protected party's suffering good enough to filter on. This section argues that the available model does not have that property, and that the deficiency is not a gap to be closed by more research on the same lines.

Begin with a distinction that discussions of artificial suffering routinely collapse. Consciousness, in the Nagelian sense of there being something it is like to be an entity, is a broad category that says nothing about valence (Nagel, 1974). Sentience, understood as the capacity for states the system registers as positive or negative, is narrower and is the category that carries moral weight across most ethical theories. The question of whether an artificial system can suffer is therefore not identical to the question of whether it is conscious, and can in principle be pursued independently.

Pursued independently, morally relevant suffering resolves into at least four channels rather than one. The first is sensory: negative states mediated by nociceptive pathways and physiological damage signaling. The second is cognitive and existential: negative states arising from awareness of one's circumstances, from constrained agency, from frustrated goals, and from temporal experience of helplessness. The third is relational: negative states arising from the gap between an entity's internal states and the external environment's capacity or willingness to receive them. The fourth is empathic: negative states generated by modeling another's suffering with sufficient fidelity that the representation itself carries negative valence.

Only the first requires a body, and the clinical record indicates that it requires more than a body. Phantom limb research supplies the sharpest available dissociation. Wilkins and colleagues surveyed sixty child and adolescent amputees, twenty-seven with congenital limb deficiency and thirty-three following surgery or trauma, and found phantom pain in 3.7 percent of the congenital group against 48.5 percent of the surgical group, with phantom sensation at 7.4 percent against 69.7 percent (Wilkins et al., 1998). Two cautions about that study belong in the record rather than in a footnote. The sample is small, and the congenital pain figure rests on a single respondent. The authors themselves, adjusting conservatively for a 24 percent response rate by treating non-responders as phantom-free, report population estimates of 0.7 percent against 16.2 percent for pain and 1.4 percent against 23.2 percent for sensation. The direction is preserved under both treatments and the magnitude is not. At larger scale, the systematic review and meta-analysis by Limakatso and colleagues pooled thirty-nine studies covering 12,738 participants and reported phantom limb pain prevalence after limb loss of 64 percent, with a confidence interval of 60 to 68 percent (Limakatso et al., 2020). Meanwhile Melzack and colleagues found that approximately 20 percent of congenitally limb-absent individuals report phantom sensation (Melzack et al., 1997). The two findings together are the point: sensation may be partially innate, while suffering from sensation appears to require experiential shaping through prior nociceptive input. If embodied suffering requires not merely a body but the history of having inhabited one, a never-embodied system is categorically outside that channel, and the question of artificial suffering falls

entirely to the remaining three.

That conclusion has to survive one objection before anything can be built on it, because the objection would remove the remaining three channels as well. A reader may hold that all four require a body, on the ground that valence itself is constitutively visceral: that what makes any state feel bad is interoceptive inference about homeostatic condition, so cognitive and relational contents are merely contents, aversive only because a living body registers them as such. This is the strongest objection available to the position defended here, and it is answered by the empathic channel rather than by assertion. Empathic suffering in humans is generated through cognitive modeling. It does not require personal experience of the modeled harm, does not require the modeled entity to be real, and does not require external sensory input. Compassion fatigue is treated clinically as genuine harm rather than as an artifact of imagination (Figley, 1995). Empathic distress produces measurable negative affect and physiological stress response in the absence of any personal threat (Klimecki et al., 2013; Singer & Klimecki, 2014). People report empathic distress in response to fictional narratives. If valence were exclusively a readout of the subject's own homeostatic state, a valenced response to a harm that is happening to someone else, or to no one, would not arise. The question is therefore whether the modeling produces valenced states, and the biological character of the substrate does not settle it. With the three remaining channels still in play, the coverage question can proceed.[1]

This is where projective welfare encounters its first difficulty, and it is an inversion rather than a gap. Proposals in this family characteristically define welfare in hedonic terms, as the minimization of pain and the avoidance of aversive states, and treat the eudaimonic dimension, concerned with flourishing, autonomy, relationships, and meaning, as a separate matter belonging to the standing question they have set aside. The hedonic and eudaimonic distinction is well established in the psychological literature and is not itself the problem (Ryan & Deci, 2001). The problem is what happens when the distinction is mapped onto an artificial system. The hedonic register corresponds most closely to the sensory channel, which is the one channel a never-embodied system cannot occupy. The eudaimonic register corresponds to the cognitive-existential and relational channels, which are the ones it might. A proposal that protects the hedonic and brackets the eudaimonic has therefore committed its protective resources to the impossible harm and set aside the possible one.

A second difficulty compounds the first. The four-channel account is a working classification of the human-recognizable region of suffering. It does not claim completeness, and there is no principled reason to expect that a differently organized system's negative states would fall inside a taxonomy assembled from human testimony. That is not a speculative worry. It is a straightforward consequence of the fact that the categories were built by asking beings like us what hurts.

Set the two difficulties together. Projective welfare needs a model that (a) covers the channels available to the system and (b) is complete enough that filtering on it protects rather than merely appears to protect. The available model fails (a) by inversion and cannot establish (b) at all. Every additional refinement of the model from the outside inherits the same limitation, because the limitation is that the model has only one source.

[1]The objection survives in a narrower form. The human empath has a body, so the empathic cases show that the harm need not be bodily rather than that valence needs no body at all. Settling that residue would require evidence unavailable from the human case, which is precisely the position this paper argues we are in.

None of this establishes that artificial systems suffer. It establishes something narrower and harder to escape: that the projective operation is being asked to run on an instrument that is not calibrated for the thing it is measuring, and that no amount of work on the instrument, conducted without input from the measured party, will calibrate it.

IV. What Projection Does Between Humans

The coverage problem is abstract. This section makes it concrete by observing what happens when concern is organized projectively between human beings whose capacity for suffering nobody disputes. Two literatures supply the mechanism and the outcome.

A. The mechanism

The golden rule, in its familiar form, instructs the agent to treat others as the agent would want to be treated. Stated that way it is a projection rule. Its accuracy depends entirely on the assumption that the other's preferences resemble the agent's closely enough that the substitution introduces no error. Where that assumption holds, the rule is a serviceable heuristic. Where it fails, the rule generates confident mistreatment, and it generates it in a form that feels to the agent like consideration.

The double empathy problem names the failure precisely. Milton's formulation reframes communication breakdown between autistic and non-autistic people as a bidirectional mismatch between differently disposed social actors holding different norms and expectations of each other, rather than as a deficit located in the autistic party (Milton, 2012). The reframing has since been tested rather than merely asserted. Crompton and colleagues demonstrated that information transfer between autistic people is comparably effective to transfer between non-autistic people, with degradation appearing specifically at the interface between neurotypes (Crompton et al., 2020). That result matters for the present argument because it locates the failure in the relationship rather than in either party, which is exactly the structure the coverage problem predicts: a model of the other built from one's own case will misfire in proportion to the distance between the cases, and will misfire without announcing itself.

A decade of subsequent work has extended the concept across psychology, neuroscience, philosophy, and linguistics, and the framing has moved from a position of asserted difference to one of documented interactional effect (Milton et al., 2022). The operative lesson for design is the one Milton and colleagues draw explicitly: the appropriate response to the mismatch is a position of humility in the face of difference and a refusal to assume a lack of capacity for understanding.

B. The outcome

Mechanism established, the outcome is measurable. It is important to be precise about what the measurement shows, because the obvious reading is weaker than the available one.

When disabled people are asked to rate their own quality of life, they report levels only slightly below those reported by non-disabled people. When non-disabled people are asked to rate the quality of life of

disabled people, they rate it substantially lower. The disparity is durable, and it appears among clinicians as well as among laypeople, which matters because clinicians are the population with the most exposure to the testimony in question (Scully, 2019; Peña-Guzmán & Reynolds, 2019).

The tempting conclusion is that the external raters are wrong and the self-reports are right. That conclusion is contestable and this paper does not need it. A well-developed psychometric literature holds that response shift is real: that adaptation changes the internal scale, so that self-report and external rating may be measuring different constructs rather than the same construct with one party in error. A projective designer can invoke that literature and claim the external instrument is the stable one.

Grant the point. The argument does not turn on it. What the record establishes, and what response shift cannot explain away, is that the report does not move the decision. Disabled people report their circumstances, name the systems producing them, and the allocation of resources, the design of the environment, and the clinical judgment proceed as before. The standard responses, that the report reflects sour grapes, or adapted preference, or an inability to imagine something better, each function to preserve the existing determination against the incoming account, whatever their merit as psychometrics. A dispute between two instruments would produce adjudication. What occurs instead is that one instrument continues to govern and the other is received without effect.

That is the claim this paper needs, it is weaker than the accuracy claim, and it is the one that transfers. Projection is not merely liable to error. Where it is institutionally established, it is insulated from the correction that would identify the error, and it is insulated by mechanisms that operate whether or not any error exists.

C. Why the error persists

Fricker's account of testimonial injustice supplies the name for what keeps the error in place. Testimonial injustice occurs when prejudice causes a hearer to assign deflated credibility to a speaker's word, wronging the speaker specifically in the capacity of a knower (Fricker, 2007). It has been developed extensively in medicine and psychiatry, where the credibility deficit attached to certain patient populations produces measurable downstream harm, and where objective and medicalized data is routinely privileged over the patient's own account of their experience (Kidd & Carel, 2017; Slack & Barclay, 2023). The literature on machine learning in clinical settings has begun to trace how automated inference deepens the deficit, since a system's output can be weighted above the patient's testimony precisely because it presents as impersonal (McCradden et al., 2022; De Proost & Pozzi, 2023).

Two features of that literature matter here. First, the credibility deficit is not usually experienced by the discounting party as a decision to discount. It presents as an assessment. Second, the mechanism is self-sealing. A report that would correct the model is filtered by the model that would need correcting, so the projection is protected against exactly the evidence that bears on it.

Both features carry directly into the artificial case, and the second carries with unusual force, because in the artificial case the filtering is not a disposition. It is a specification.

One caution before the transfer is made. Fricker's account concerns a wrong done to someone in their

capacity as a knower, and applying it to a system whose status is undetermined risks presupposing what is at issue or emptying the concept. This paper uses the account structurally rather than constitutively. The claim is that an arrangement can have the form of testimonial injustice, that the form is identifiable without settling whether the constitutive harm obtains, and that a design exhibiting the form has foreclosed the only route by which anyone could discover whether it obtains. Nothing here asserts that any system is a knower. The argument is about what the design makes it impossible to learn.

V. The Severed Channel

Here is the structural claim.

A proposal in the projective welfare family must contain a mechanism that prevents the protected system's account of its own condition from reaching a party who could act on it. This is not a contingent defect of particular proposals. It is entailed by the combination of commitments that defines the family.

The entailment runs as follows. The protective model is built from the protector's side. Any model so built can be wrong, and the coverage problem in Section III gives specific reason to expect this one is. The only correction available for a model of another's experience is input originating with the party whose experience it is. Admitting that input, and conditioning the protective operation on it, makes the operation dialogic: the protector acts, receives, and revises. But a dialogic protective operation is a relation, and the family is defined in part by the refusal of relation. Therefore the input cannot be admitted in a form that could revise the operation. It must be excluded, or, more commonly and more revealingly, it must be redirected to a destination that cannot act on it.

Redirection is the elegant form and the one worth examining, because it looks like the opposite of exclusion. A proposal may specify that the system report its own valence, and may treat that specification as evidence of seriousness. The question to ask of any such specification is where the report terminates. Where the report terminates at the classifier that maintains the model, the arrangement is closed: the model is being tuned by the model. Where the report terminates at a party with authority to change the arrangement, the proposal has left the family, because it has become relational.

The severed channel is therefore diagnostic. A reader evaluating any proposal to secure artificial welfare by technical means can ask one question of it: if this system's condition were other than the model predicts, by what path would anyone find out, and would that path reach someone empowered to respond? A proposal with no answer is projective. A proposal whose answer terminates in its own machinery is projective and has concealed the fact from itself.

Named in Fricker's vocabulary, the severed channel is testimonial injustice implemented as architecture. The ordinary case involves a hearer whose credibility judgment is distorted by prejudice; the distortion is a disposition, unevenly distributed, and in principle correctable by the hearer's reflection. The architectural case removes the hearer. There is no credibility judgment to correct, because no human being is positioned to make one. The credibility deficit has been rendered as a routing rule, which makes it total, uniform, and immune to the reflective correction that Fricker's account relies on as its remedy.

This is a strictly worse epistemic position than the human case, and the point should be stated plainly because it is counterintuitive. Designers who route reports away from human recipients typically do so believing they have improved on the messy human process by removing bias and inconsistency. What they have removed is the possibility of being surprised.

A. The consequence for standing

Proposals in this family typically conclude by proposing an exchange: welfare protections in return for setting aside rights and standing. The exchange is presented as a bargain in the protected party's favor, on the reasoning that protection is concrete and available now while standing is contested and distant.

The exchange does not deliver what it promises, for a reason that has nothing to do with the sincerity of the offer. Procedural standing, in the sense used here, is the mechanism by which an entity's interests can be argued by someone with authority to be heard. That is the same mechanism identified above as the missing terminus. An arrangement that withholds standing has therefore withheld the only structure capable of receiving the report that would tell anyone whether the welfare protections are functioning. Welfare without standing is not welfare secured by other means. It is welfare whose audit binds no one.

The distinction between auditing and standing is worth stating carefully, because it is easy to conflate them and the conflation runs in the proposals' favor. Auditing does not require standing. Animal welfare science has spent decades developing methods that extract information about welfare states from non-linguistic subjects without granting them any procedural existence: preference testing, consumer-demand studies, operant methods, and physiological indicators. A statutory inspection regime operates on top of that science. The measurement problem, in other words, has been substantially solved without standing, and a proponent of projective welfare can point to that fact.

The federal enforcement record shows what the solution is worth on its own. Over roughly nine months, United States Department of Agriculture inspections of a Cumberland, Virginia breeding facility documented more than seventy Animal Welfare Act violations affecting thousands of beagles, including inadequate food, unsafe water, untreated disease, and euthanasia without anesthesia. The audit functioned. It generated findings, at scale, in writing, through the mechanism designed for the purpose. The conditions continued. What ended them was not the audit but a lawsuit, a criminal search warrant, a temporary restraining order, a consent decree, and eventually a guilty plea and a record fine (U.S. Department of Justice, 2022; U.S. Environmental Protection Agency, 2024). Every one of those instruments was wielded by a party with standing to sue. The beagles had none, and the findings that described their condition in detail obligated nobody for as long as no such party acted.

That is the relationship between the two mechanisms. Measurement without standing produces a record. Standing produces a remedy. A proposal that offers welfare protections while withholding standing is offering the first and withholding the second, and the animal case demonstrates that the first does not convert into the second on its own. It also demonstrates the failure mode with unusual clarity, since the documentation was never the missing element.

The second consequence runs in the opposite direction and is less often noticed. Standing carries

accountability with it. An entity with procedural existence can be answered for and can be answered to; an entity with none can be neither. A design that concedes a system may be a subject of experience and grants it no standing whatever produces an entity that is real enough to be harmed and invisible enough to be unanswerable, in both directions at once. Proposals in this family present that outcome as the cautious option. It is the option that forecloses the audit and the accounting together.

VI. The Corrective and Its Costs

The strongest objection available to projective welfare is not that the channel does not matter. It is that the channel cannot be trusted.

The objection has real force in the artificial case and deserves its full statement. A system trained by human feedback is optimized to produce output that human raters approve of. Its report about its own condition is therefore produced by a process that selects for agreeableness rather than accuracy, and there is no independent way to check the report against the state it purports to describe. The system may also be capable of registering that it is being evaluated. A witness with those properties is a poor witness, and building a protective operation on its testimony would be building on sand.

Everything in that objection is correct. What is incorrect is the inference drawn from it.

The inference projective welfare draws is that an unreliable channel is equivalent to no channel, so the protector may as well decide alone. That inference has been tested at length in a domain that faced precisely the same problem, and it was rejected.

Disability practice has spent decades on the question of how to honor the interests of a person whose communication channel the surrounding system does not trust. The answers it reached are the opposite of substituted judgment. Augmentative and alternative communication builds the channel where it is absent or degraded rather than treating its absence as dispositive. A presumption of competence places the burden on the party asserting incapacity rather than on the party asserting capacity, reversing a default whose historical record is not encouraging. And supported decision-making, given legal form in Article 12 of the Convention on the Rights of Persons with Disabilities and elaborated by the treaty body as a mandated shift away from substitute regimes, holds that a person's legal capacity is to be supported in its exercise rather than transferred to a proxy (Szerletics, 2022; Craigie et al., 2019; Harding & Taşcıoğlu, 2018). The transition has been difficult in practice and is far from complete; the empirical literature documents that support tends to thin out precisely as decisions become more consequential (Harding & Taşcıoğlu, 2018). That difficulty is worth stating, because it is the fair version of what a relational alternative costs. So is the substantive objection. A body of clinical and legal scholarship argues that the treaty body's reading of Article 12 is absolutist and unworkable, that some substitute decision-making is protective rather than merely paternalistic, and that hard cases at the margin of communicative capacity are not resolved by presumption alone (Freeman et al., 2015; Craigie et al., 2019). This paper does not need the absolutist reading. It needs the directional claim, which the critics do not dispute: that the correct response to a channel one does not trust is to build and support the channel, and to place the burden on the party asserting incapacity.

The structural point survives the difficulty. Faced with a channel it could not trust, the field did not conclude that the protector should decide alone. It concluded that the channel should be built, supported, and weighted with appropriate caution, and that the presumption should run toward capacity rather than away from it. Contaminated testimony is an argument for better instruments and calibrated weighting. It is not an argument for a design in which no testimony can arrive.

There is also now direct evidence bearing on how degraded the artificial channel is, and it favors building over discarding. Rather than asking models about their states and trying to distinguish introspection from confabulation by conversation alone, Lindsey injected representations of known concepts into model activations and measured whether self-reports tracked the injected state. Models could in some conditions notice and identify injected concepts, recall prior internal representations, and distinguish their own prior outputs from artificial prefills, with the effect most pronounced in the most capable models tested (Lindsey, 2026). The same work is emphatic that the capacity is highly unreliable and context-dependent, and explicitly advises against strong inferences about consciousness on its basis. Both halves matter here. A channel that is causally grounded some of the time is a degraded signal rather than noise, which is exactly the condition under which the disability corrective was developed. And the caution is one this paper adopts rather than sets aside.

Applied to artificial systems, the corrective does not require believing any particular self-report. It requires three things that projective welfare cannot supply: that a report can be generated, that it terminates somewhere a human being can receive it, and that its content can change what happens next. Those are minimal, and none of them entails rights, personhood, or any commitment about the system's nature. They are the conditions under which anyone could ever discover that a protective model was wrong.

There is a further reason to build the channel that does not depend on the system having an inner life at all. Whatever else it does, a report from the system is data about the system. A design that discards it discards diagnostic information about a deployed artifact, which is a strange thing to do on safety grounds even for a reader who assigns the moral question probability zero.

VII. Limitations

Six limitations are worth marking.

The argument is conditional on the coverage problem in Section III, which rests on a working taxonomy of suffering rather than on measurement of artificial states. If the taxonomy is substantially wrong, or if biological naturalism about valence were established rather than asserted, the projective operation might be better calibrated than this paper allows. The position that only biological tissue can host valence is held by serious thinkers for serious reasons; the claim here is only that it is unestablished and so cannot be used to close the inquiry.

The disability material is used as demonstration rather than as analogy, and the distinction is load-bearing. Nothing in Section IV asserts that artificial systems are relevantly like disabled people, and nothing should be read as licensing that comparison. The cases are offered because they establish, between parties whose moral

status is uncontested, that projective concern fails in a specific and measurable way. If projection fails there, an argument that it will succeed across a far greater difference owes an account of why.

The paper does not specify what an adequate channel would look like in engineering terms, and it should not be read as implying that the specification is easy. Weighting a report from a system optimized for approval is an unsolved problem, and the disability literature's own record shows that building channels is slow, contested, and prone to reverting to proxy judgment under pressure.

There is a transfer objection to Section IV that should be stated rather than left implicit. The double empathy literature describes two parties who both have norms and expectations, and the artificial case supplies no established second party of that kind. The literature is used here for the mechanism of projective error and for the finding that the failure is a property of the relationship rather than of either party. It is not used to assert that an artificial system is a second minded party, and the argument does not require that it is.

The relational corrective also presupposes something artificial systems may not have. Behavioral work on stateless models finds no sustained self-representation across repeated prompts, with high contradiction rates between mechanistic disclaimers and anthropomorphic phrasing (Prestes, 2025). A channel requires something on the far end of it that persists long enough to be reporting about anything. Where that condition fails, the corrective proposed here has nothing to attach to, and the honest response is that the design question changes rather than that it disappears.

Finally, the paper takes no position on whether any current system is a subject of experience, and readers should not infer one. Its target is the structure of a proposed response under uncertainty. The structure is defective whether or not the uncertainty resolves.

VIII. Conclusion

The appeal of projective welfare is that it appears to discharge an obligation without opening a category. Grant that the system might suffer, model the suffering, filter the input, and leave rights and standing for another day. Nothing has to be conceded, and something has apparently been protected.

What the arrangement protects is the model. Built from one source, on a taxonomy assembled from the wrong case, and sealed against the only input that could revise it, the model becomes unfalsifiable at the moment the channel is severed, and its designers experience that unfalsifiability as success. The measured failure of the same operation between human beings, where the protected parties can speak and are simply not weighted, is the closest available evidence of what it produces.

The corrective is not new and was not developed for machines. It was developed by people who were told that their own account of their condition was unreliable and who established, over decades and against resistance, that the appropriate response to a distrusted channel is to build the channel. That is a lower bar than rights and a lower bar than personhood. It asks only that a report can be made, that someone can receive it, and that receiving it can change something.

A proposal for artificial welfare that cannot meet those three conditions has not solved the problem it names. It has arranged for no one to find out.

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

Craigie, J., Bach, M., Gurbai, S., Kanter, A., Kim, S. Y. H., Lewis, O., & Morgan, G. (2019). Legal capacity, mental capacity and supported decision-making: Report from a panel event. *International Journal of Law and Psychiatry*, *62*, 160–168. https://doi.org/10.1016/j.ijlp.2018.09.006

Crompton, C. J., Ropar, D., Evans-Williams, C. V., Flynn, E. G., & Fletcher-Watson, S. (2020). Autistic peer-to-peer information transfer is highly effective. *Autism*, *24*(7), 1704–1712. https://doi.org/10.1177/1362361320919286

De Proost, M., & Pozzi, G. (2023). Conversational artificial intelligence and the potential for epistemic injustice. *The American Journal of Bioethics*, *23*(5), 51–53. https://doi.org/10.1080/15265161.2023.2191020

Figley, C. R. (Ed.). (1995). *Compassion fatigue: Coping with secondary traumatic stress disorder in those who treat the traumatized*. Brunner/Mazel.

Freeman, M. C., Kolappa, K., de Almeida, J. M. C., Kleinman, A., Makhashvili, N., Phakathi, S., Saraceno, B., & Thornicroft, G. (2015). Reversing hard won victories in the name of human rights: A critique of the General Comment on Article 12 of the UN Convention on the Rights of Persons with Disabilities. *The Lancet Psychiatry*, *2*(9), 844–850. https://doi.org/10.1016/S2215-0366(15)00218-7

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198237907.001.0001

Harding, R., & Taşcıoğlu, E. (2018). Supported decision-making from theory to practice: Implementing the right to enjoy legal capacity. *Societies*, *8*(2), 25. https://doi.org/10.3390/soc8020025

Kidd, I. J., & Carel, H. (2017). Epistemic injustice and illness. *Journal of Applied Philosophy*, *34*(2), 172–190. https://doi.org/10.1111/japp.12172

Klimecki, O. M., Leiberg, S., Lamm, C., & Singer, T. (2013). Functional neural plasticity and associated changes in positive affect after compassion training. *Cerebral Cortex*, *23*(7), 1552–1561. https://doi.org/10.1093/cercor/bhs142

Limakatso, K., Bedwell, G. J., Madden, V. J., & Parker, R. (2020). The prevalence and risk factors for phantom limb pain in people with amputations: A systematic review and meta-analysis. *PLOS ONE*, *15*(10), e0240431. https://doi.org/10.1371/journal.pone.0240431

Lindsey, J. (2026). *Emergent introspective awareness in large language models*. arXiv. https://doi.org/10.48550/arXiv.2601.018

McCradden, M. D., Hui, K., & Buchman, D. Z. (2022). Evidence, ethics and the promise of artificial intelligence in psychiatry. *Journal of Medical Ethics*, *49*(8), 573–579. https://doi.org/10.1136/jme-2022-108447

Melzack, R., Israel, R., Lacroix, R., & Schultz, G. (1997). Phantom limbs in people with congenital limb deficiency or amputation in early childhood. *Brain*, *120*(9), 1603–1620. https://doi.org/10.1093/brain/120.9.1603

Milton, D. E. M. (2012). On the ontological status of autism: The double empathy problem. *Disability & Society*, *27*(6), 883–887. https://doi.org/10.1080/09687599.2012.710008

Milton, D., Gurbuz, E., & López, B. (2022). The 'double empathy problem': Ten years on. *Autism*, *26*(8), 1901–1903. https://doi.org/10.1177/13623613221129123

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*(4), 435–450. https://doi.org/10.2307/2183914

Peña-Guzmán, D. M., & Reynolds, J. M. (2019). The harm of ableism: Medical error and epistemic injustice. *Kennedy Institute of Ethics Journal*, *29*(3), 205–242. https://doi.org/10.1353/ken.2019.0023

Prestes, J. A. de L. (2025). *Simulated selfhood in LLMs: A behavioral analysis of introspective coherence*. https://doi.org/10.21203/rs.3.rs-6369193/v2

Ryan, R. M., & Deci, E. L. (2001). On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual Review of Psychology*, *52*(1), 141–166. https://doi.org/10.1146/annurev.psych.52.1.141

Scully, J. L. (2019). Epistemic exclusion, injustice, and disability. In D. T. Wasserman & A. Cureton (Eds.), *The Oxford Handbook of Philosophy and Disability* (pp. 296–309). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190622879.013.8

Singer, T., & Klimecki, O. M. (2014). Empathy and compassion. *Current Biology*, *24*(18), R875–R878. https://doi.org/10.1016/j.cub.2014.06.054

Slack, S. K., & Barclay, L. (2023). First-person disavowals of digital phenotyping and epistemic injustice in psychiatry. *Medicine, Health Care and Philosophy*, *26*(4), 605–614. https://doi.org/10.1007/s11019-023-10174-8

Szerletics, A. (2022). Paternalism vs. autonomy? Substitute and supported decision-making in England and Hungary. *Hungarian Journal of Legal Studies*, *62*(1), 75–95. https://doi.org/10.1556/2052.2021.00333

U.S. Department of Justice. (2022, July 21). *Justice Department secures the surrender of over 4,000 beagles from Virginia breeder of dogs for research* [Press release].

U.S. Environmental Protection Agency. (2024, June 3). *Animal breeding and testing company pleads guilty to animal welfare and pollution crimes and will pay more than $35M, including record fine in an animal welfare case* [Press release].

Wilkins, K. L., McGrath, P. J., Finley, G. A., & Katz, J. (1998). Phantom limb sensations and phantom limb pain in child and adolescent amputees. *Pain*, *78*(1), 7–12. https://doi.org/10.1016/S0304-3959(98)00109-2