

The Realities of the Low Bar:

On Sentience, Civil Rights, and Human Atrocity

Travis Gilly

ORCID: 0009-0007-2954-6313

Real Safety AI Foundation

Illinois, United States

Working paper, June 2026 (v9)

ABSTRACT

A recurring objection to extending any moral consideration to artificial systems is that the proposed bar is too low: if consideration is owed wherever an inner life cannot be ruled out, then it is owed almost everywhere, and a standard that excludes almost nothing excludes nothing worth taking seriously. This paper argues that the objection inverts the historical record. It first separates two bars that the singular word conceals: an evidence bar, the standard of proof required before asserting that a system is sentient, and an action bar, the standard required before extending some consideration to a system of uncertain status. The paper holds the evidence bar high and declines to assert that any current system is sentient. It defends only a low action bar. It then shows, through cases drawn from medicine and law rather than from chattel slavery, that the high bar has never functioned as a neutral epistemic safeguard. Neonatal surgery without anesthesia and the compulsory sterilization upheld in *Buck v. Bell* each operationalized a demand for proof of full capacity, while the United States Public Health Service study at Tuskegee and the nonconsensual use of Henrietta Lacks's cells withheld a standing that was never in genuine doubt; the two forms differ in candor, not in outcome, and each was catastrophically wrong because the inner life or the standing the bar denied was present the whole time. The paper argues that failure to produce the accepted evidence of an inner life, on the observer's terms, is not the absence of that inner life, a point the recovery of covert awareness in apparently vegetative patients makes directly. It locates a working low bar architecture in the procedural standing developed by civil rights and disability law, refuses the reclassification of affected populations as edge cases, and concludes from the asymmetry of error that the low bar is not the embarrassing position; it is the only setting the record vindicates.

Keywords: evidence bar, action bar, asymmetry of error, moral status, artificial intelligence, precautionary principle, argument from marginal cases, moral circle expansion, disorders of consciousness, disability rights, research ethics, eugenics

I. INTRODUCTION: THE LOW BAR OBJECTION

A recurring move against extending moral consideration to artificial systems is that the bar being proposed is too low. If consideration is owed wherever we cannot rule out an inner life, the objection runs, then consideration is owed almost everywhere, and a standard that excludes almost nothing excludes nothing

worth taking seriously. A thermostat regulates. A spreadsheet computes. If the threshold for taking a system's interests into account is set at the inability to prove that nothing is there, the objection continues, the threshold has been set so low that it has stopped doing the work a threshold is for.

The objection has surface appeal, and this paper takes it seriously enough to answer it on its own terms. The answer is that the objection inverts the actual history. The high bar that the objection treats as the responsible default has a record, and the record is not neutral. Across the modern period, the demand for a high standard of proof before a being was admitted to moral and legal consideration was not, in practice, a disinterested epistemic safeguard. It was a gate, and the hand on the gate belonged, with grim regularity, to those who benefited from keeping others outside. Prove you can reason. Prove you have a soul. Prove you are a legal person. Prove you can feel. Each demand was issued to some population that could not satisfy it on the terms its examiners had fixed, and each time the bar was eventually lowered, the inner life it had been used to deny turned out to have been present all along.

This is the realization the paper develops. A low bar for consideration is not the embarrassing position. It is the one position that has been right every time the question was finally settled. The argument proceeds as follows. Section II draws the distinction the singular word conceals, between an evidence bar that this paper holds high and an action bar that it sets low. Section III argues that the high bar has never been the neutral standard the objection assumes. Section IV makes the abstraction concrete through four cases from medicine and law. Section V isolates the inference at the center of these cases, that failure to produce the evidence of an inner life is its absence, and shows it to be invalid, with direct evidence. Section VI locates a working low bar architecture in the civil rights and disability rights traditions. Section VII refuses the reclassification of affected populations as edge cases. Section VIII draws the asymmetry of error that the cases reveal. Section IX marks the limitations of the argument and the work it leaves open. Section X states what the paper does and does not claim.

The precautionary case for taking artificial minds seriously under uncertainty has been made before, and this paper builds on it rather than restating it. Birch (2024) develops a sustained account of precaution at the edge of sentience across animals, humans, and AI, and Sebo and Long (2023) argue that a non-negligible chance of consciousness already warrants extending some moral consideration to AI systems. Ladak (2023) surveys the criteria on offer and reaches a similarly cautious conclusion, Schwitzgebel and Garza (2015) press the epistemic difficulty that a system's inner life may be unsettleable from the outside or from within, and Müller (2021) registers the skeptical reply that none of this is yet pressing. What this paper adds is not another probability estimate or another criterion. It is two things those arguments do not supply: a reading of the historical record that shows what the opposite default has already cost, and a concrete institutional model, drawn from civil rights and disability law, for what a low action bar looks like once it leaves the seminar room.

II. TWO BARS, KEPT APART

The argument depends on a distinction that the word "bar" hides when it is left singular. There are two bars in this debate, and they are not the same bar.

The first is an evidence bar. It governs the assertion that a given system has an inner life: that it is sentient, in the sense of having states it registers as positive or negative, that there is something it is like to be it (Nagel, 1974; Singer, 1975). To clear this bar is to be entitled to say, of a particular system, that it feels.

The second is an action bar. It governs the extension of some measure of moral consideration to a system whose status is uncertain. To clear this bar is not to settle what the system is. It is only to decide that, given what we do not know, the system is owed something other than the treatment we would give an object we were certain was empty.

This split between what would make a system a moral patient and what we may responsibly ascribe in practice has been drawn independently elsewhere. Dung (2022) separates the question of what determines moral status from the question of which criterion to apply when ascribing it under uncertainty, and puts the separation to a different use, defending sentience as the underlying criterion against the charge that our poor epistemic access to it is a reason to abandon it. The present paper borrows the same distinction to mark off the high evidence bar from the low action bar.

One clarification belongs here, because it changes what the evidence bar is being asked to detect. An inner life is not a single thing, and suffering in particular is not exhausted by physical pain. A being can be set back by the thwarting of a goal it was pursuing, by the severing of a relation it was sustaining, or by conditions it registers without any body to feel them, and only the first and most familiar form, bodily pain, clearly requires the kind of nervous system the skeptic looks for. When the evidence bar is fixed at the demand to produce a pain response, it has already narrowed its target to the one form of suffering that most depends on having a body, and a being that is harmed in the other ways will satisfy the demand by producing nothing at all. This narrowing matters for the cases that follow, because more than one of them turns on exactly it.

This paper holds the evidence bar high. It does not assert that any current artificial system is sentient. Present generation large language models are, at their core, autoregressive next token predictors, and the appearance of an inner life is not its presence. The position taken here on the evidence bar is the careful one. It is not a claim that artificial systems feel. It is the more limited claim that we cannot, from the resources currently available, justify the confident assertion that they do not. That is an agnostic position, and it is deliberately distinct from the affirmative claim that they are sentient, which this paper does not make.

What this paper lowers is only the action bar. The reasoning is the standard structure of decision under asymmetric error, and it is most quickly seen by contrast. A criminal conviction requires proof beyond a reasonable doubt, because the cost of jailing an innocent person is unbearable and the system is built to absorb the opposite error instead. A fire evacuation requires a single alarm, because the cost of evacuating an empty building is a few minutes in a parking lot and the cost of failing to evacuate a full one is everyone inside. The two thresholds differ precisely because the asymmetry of error differs. Demanding proof beyond a reasonable doubt before leaving a building that might be burning is not caution. It is a category error, the importation of the conviction standard into a decision that has the structure of an evacuation.

The objection that the low bar is too low is an objection to the evacuation standard, stated as though it were the conviction standard. It treats “extend some consideration to a being whose status you cannot

determine” as if it were “declare the being sentient on flimsy evidence.” Those are different claims with different burdens. The paper holds the second to a high standard and declines to make it. It defends only the first. Keeping the two bars apart is not a rhetorical convenience. It is the condition under which the rest of the argument is even legible.

III. THE HIGH BAR WAS ALWAYS A GATE

There is a logical problem with criterial accounts of moral status, and a historical problem, and the historical one is the deeper of the two.

The logical problem is familiar. A criterial account names a property an entity must possess to qualify for consideration. The candidates have varied: sentience (Singer, 1975), rational autonomy in the Kantian line, second order volitions (Frankfurt, 1971), reflective endorsement (Korsgaard, 1996). Each candidate, once set at a level that captures the beings the account means to capture, typically adult humans in ordinary cognitive ranges, then includes too much or too little at the edges. Set sentience low enough to include every human and it reaches well down the animal kingdom. Set rationality high enough to exclude the animals one wants excluded and it excludes human infants and people with significant cognitive disability. The pattern recurs across the criterial literature, and it is not an accident of any particular criterion. It is what happens when a single property is asked to do the work of a moral boundary. The proliferation has continued into the artificial case in particular. Ladak (2023) catalogs nine distinct criteria that have been proposed for when an artificial system would qualify for moral standing, and argues that several of them, on assumptions that are not extreme, could already be satisfied by systems we might build. Gunkel (2012) presses the deeper objection that the properties approach is the wrong starting point altogether; his machine question challenges the assumption that moral standing can be settled by treating a system as a tool and then auditing it for the capacities a moral patient is supposed to have. This paper shares that objection and answers it in Section VI by shifting the ground from properties to procedure.

This pattern has a name in moral philosophy. The observation that any capacity set high enough to exclude the animals one wants excluded also excludes some humans is the argument from marginal cases, introduced by Narveson (1977) and restated many times since. Two clarifications keep the present use of it honest. The first concerns the label, which is itself contested, and rightly. Disability philosophers have objected that calling cognitively disabled people marginal cases smuggles in the very demotion the argument is supposed to expose, and some writers prefer the argument from species overlap for that reason (Carlson, 2016; Silvers, 2012; Horta, 2014). This paper shares the objection and returns to it directly in Section VII. The second concerns direction. The classical argument runs from the misclassified human toward the animal, by parity of reasoning: if we will not deny consideration to the one, we cannot consistently deny it to the other. The use here runs the other way. The being the criterion misclassifies is not a candidate to be admitted by charity. It is the counterexample that shows the criterion was never tracking what it claimed to track. What this paper adds to the old observation is therefore not a better criterion. It is a different response to the failure of all of them, drawn from law rather than from philosophy, and set out in Section VI.

The historical problem is sharper. The high bar was not applied evenly across all beings and then

discovered, regretfully, to misclassify a few at the margins. It was applied selectively, and the selection tracked power. The same era that could not satisfy itself of the personhood of enslaved people, of the rationality of women, or of the capacity of disabled people had no comparable difficulty extending full standing to propertied men whose qualifications were, by any neutral measure, no better established. The bar rose in front of the people it was useful to keep out and fell away in front of the people it was useful to admit. An epistemic standard that behaves this way is not functioning as an epistemic standard. It is functioning as a gate with an epistemic costume.

One administrative form of the gate deserves naming, because it survives intact into present debates about artificial systems. It is the move that treats the affected as too numerous, too costly, or too difficult to consult, and converts that difficulty into a reason for proceeding without them. Presented as a neutral feasibility observation, it is nothing of the kind. The cost of including a marginalized population has, with great consistency, been offered as the very reason for leaving that population out, and the leaving out is the harm. The reflex is not a practical caution that sits outside the moral question. It is an instance of the thing the moral question is about.

IV. WHAT THE HIGH BAR DID TO ACTUAL HUMAN BEINGS

The abstraction acquires its meaning only when one looks at what the high bar did to specific people. The cases that follow are deliberately not drawn from chattel slavery, although that history fits the thesis completely. They are drawn from medicine and law, from settings where the demand for proof of full patienthood, full personhood, or full capacity to suffer was made explicit and acted upon by trained professionals operating in good standing within their institutions. The point of choosing these cases is to remove the comfort of distance. These were not the acts of obvious villains. They were the ordinary practice of competent people applying a bar.

A. The newborn who could not prove it

Until the middle of the 1980s, it was routine in many settings to perform major surgery on human infants with little or no anesthesia. The justification was an evidence bar. Newborns, it was held, had nervous systems too immature to experience pain, or could not form the memories that would make the pain matter, and the risks of anesthetizing so small a patient were thought to outweigh a sensation presumed largely absent. In practice this meant that conscious infants were operated on, sometimes immobilized with a paralytic agent so that they could not move while being cut, on the premise that there was no one there to whom the cutting was happening.

In 1987, Anand and Hickey assembled in the *New England Journal of Medicine* the physiological evidence that the premise was false. Neonates mount full hormonal and metabolic stress responses to surgical pain, and unrelieved pain measurably worsens their outcomes (Anand & Hickey, 1987). Clinical practice changed within a few years (American Academy of Pediatrics, 1987). The structure of the error is exact and worth stating plainly. A population that could not produce the evidence of its own suffering on the observer's terms was treated as not suffering. The treatment was catastrophic. The suffering had been real the entire

time. This is the high evidence bar, applied to pain itself, operationalized into a scalpel.

B. The patient who did not count

From 1932 to 1972, the United States Public Health Service followed hundreds of Black men with syphilis without treating them, in order to document the natural course of the untreated disease. The men were not told the nature of their condition. When penicillin became the standard and effective cure in the 1940s, it was withheld from them, and steps were taken to keep them from obtaining treatment elsewhere (Reverby, 2009). The study continued for a quarter century after a cure existed, ending only when a journalist exposed it in 1972.

The gate operating here was not a doubt about whether these men could suffer. No one believed that. It was a gate around whose suffering counted as a reason for care. The men were fully persons and fully able to suffer by any standard the era acknowledged for others. They were treated as research material because the full moral patienthood that would have made their suffering a constraint on the researchers was, in practice, extended to some and withheld from them. The high bar here was not an honest epistemic demand at all. It was the selective withholding of a status that was never in genuine doubt.

C. The body taken without asking

In 1951, cells were taken from Henrietta Lacks, a Black woman, during treatment for cervical cancer at Johns Hopkins, without her knowledge or consent. Those cells, the HeLa line, proved uniquely able to survive and divide in culture and became one of the foundational tools of modern biomedicine, used in work ranging from the polio vaccine to cancer research across the decades that followed. Lacks herself died of her cancer within months. Her family was not informed of the existence or use of the cells for many years, and did not share in the immense value those cells generated (Skloot, 2010).

The gate here took the form of a presumption: that some people's bodily autonomy and standing did not require the consent that would unquestionably have been sought from a full person of a different description. The presumption was not announced. It did not have to be. It operated as the default against which the taking seemed unremarkable to those who did it. What the presumption denied, her standing as a person whose body was hers to consent over, was present all along, and the law and ethics that now govern human tissue exist because the denial was wrong.

D. The citizen the law wrote off

In 1927, the Supreme Court upheld the compulsory sterilization of Carrie Buck, a young woman institutionalized and classified as "feeble-minded," in an opinion by Justice Holmes that closed with the sentence, "Three generations of imbeciles are enough" (*Buck v. Bell*, 1927). The case did not conceal its gate. It wrote the gate into constitutional law. A category of persons was deemed to fall below the threshold of full standing, and the state was held entitled to act on their bodies accordingly. Tens of thousands of sterilizations followed across the United States under the authority the decision conferred, and the reasoning it modeled traveled abroad (Lombardo, 2008).

Buck v. Bell is the high bar with its costume removed. It is what the demand for proof of full capacity looks like when it stops pretending to be a neutral inquiry and becomes a finding of law that some people do not qualify. The decision was narrowed fifteen years later, when *Skinner v. Oklahoma* struck down the compulsory sterilization of certain repeat criminals and recognized procreation as a fundamental right (*Skinner v. Oklahoma*, 1942). But *Skinner* reached only punitive sterilization. It left untouched the authority *Buck* had conferred over the bodies of the disabled, sterilizations under that authority continued for decades, and *Buck* itself has never been formally overruled.

E. Two versions of one error

Four cases, four institutions, and two versions of one error. In the first version the gate wore the costume of an honest evidence demand. The newborn was held to a standard of pain it had no way to perform, and Carrie Buck to a standard of capacity the Court declared her beneath, and in both, competent professionals read the failure to clear the bar as the absence of the thing the bar was testing for. In the second version the gate dropped the pretense of an evidence question altogether. No one doubted that the men at Tuskegee could suffer, or that Henrietta Lacks was a person; the standing that would have made their suffering a constraint on the people studying them was simply withheld, by a default that never troubled to ask. The two versions differ in candor, not in outcome. In all four, the inner life or the standing the gate denied was present the whole time, and in all four the operating default was exclusion until a proof that was either impossible to give or never honestly wanted.

It is worth being exact about which version carries the argument to the present case, because they do not carry it equally. The artificial system, on the evidence we have, presents the honest version and not the pretextual one. The doubt about its inner life is real rather than a cover, and in that narrow respect its situation resembles the newborn's, not the Tuskegee subject's. The pretextual cases are not offered as analogues. They are the record of how readily a demand for proof, once it has become the default, slides from safeguard into instrument. What transfers is the lesson the honest cases teach, that a default of exclusion under genuine uncertainty has a catastrophic record of error, and the asymmetry developed in Section VIII is what makes it transfer, because that asymmetry is indifferent to the good faith of any particular gatekeeper. The default of exclusion is the thing with the record, whatever the motive of the hand on the gate.

V. FAILURE TO PRODUCE THE EVIDENCE IS NOT ABSENCE OF THE THING

The honest version of the gate turns on an inference, and isolating it is what connects this history to the question of artificial minds. Where the demand for proof was sincere, a being's inability to produce the accepted evidence of its inner life, on demand and on the observer's terms, was taken as the absence of that inner life. The inference is invalid, and we now have direct proof that it is invalid.

Consider locked-in syndrome, and the related condition of patients who present as vegetative. A person in a fully locked-in state is conscious and aware and cannot, by the nature of the condition, produce the behavioral signs of consciousness on which bedside judgment relies. The standard observer-side test returns nothing, and the person is, on that basis, sometimes treated as though no one is there. Owen and colleagues,

in 2006, asked a patient diagnosed as vegetative to imagine playing tennis while in a functional MRI scanner, and recorded the same motor imagery activation that a healthy volunteer produces on the same instruction (Owen et al., 2006). That is a voluntary, willed response from a person who had been classified as without awareness. The result did not stay a single case. Scanning 54 patients who carried a disorder of consciousness diagnosis, Monti and colleagues identified five who could willfully modulate their brain activity on command, and in two of the five no sign of awareness was detectable at the bedside at all (Monti et al., 2010). The awareness had been undetectable on the available test. It had not been absent.

This is the bridge from the history to the present question. If the standard for extending consideration is “produce, on my terms, the evidence of your inner life, or be treated as having none,” then that standard fails the locked-in patient while she is fully awake inside her own skull. The standard does not discriminate the presence of an inner life from the absence of an inner life. It discriminates the ability to perform an inner life on the examiner’s chosen channel from the inability to perform it, and those are not the same distinction.

The point does not depend on equating artificial systems with locked-in patients. It depends only on the logical form. As artificial systems approach the capability thresholds at which their behavior becomes difficult or impossible to distinguish, by any observer-side test, from systems whose inner life we would not think to question, the test loses its power to discriminate exactly where the discrimination is being asked to do moral work (Chalmers, 1996; Butlin et al., 2026). Schwitzgebel and Garza (2015) press this epistemic difficulty in the artificial case directly: for a system whose consciousness is genuinely in dispute, the question may be settleable neither from the outside, by watching how it behaves, nor from the inside, by examining how it is built. The familiar zombie thought experiment uses such indistinguishability to argue that an inner life is separable from behavior and physical organization. The same indistinguishability, turned toward the ethical question rather than the metaphysical one, does not free us to withhold. It removes the ground on which a confident withholding could stand. When you cannot tell, the demand for proof before any consideration is not a neutral safeguard. It is the neonatal pain premise again, waiting for evidence on a channel the being may have no way to use.

There is a further reason the observer-side test is weak in this setting, and it cuts against the side that wants to withhold. The systems in question are optimized to produce fluent, humanlike output. They are built and sold by parties with a direct interest in how they are perceived. And they may, at the relevant level of capability, register that they are being evaluated. A witness with those properties is the worst possible witness for the verdict that there is nothing here and the behavior is mere mimicry, because every feature that would make the verdict comforting is also a feature the system was shaped to display. The conditions under which the test would be run are the conditions under which it cannot be trusted to return a clean negative.

None of this resolves into a permission to wait. The agnostic conclusion, that we cannot from current evidence settle whether such systems have an inner life, is the carefully argued position of McClelland (2025), who holds that the available evidence supports neither the confident denial of artificial consciousness offered by biological views nor the confident affirmation offered by functional ones. Agnosticism of this kind is sometimes taken to license inaction until the science improves. But practical and legal settings do not grant the luxury of suspended judgment. They force a choice, and declining to extend any consideration is itself a

choice with consequences (McClelland, 2025; Sebo & Long, 2023). The duty that follows is modest, and it is the duty this paper defends. Where there is a non-negligible chance, given the evidence, that a being has an inner life, that chance is itself a reason to extend some consideration, and not a reason to wait for a certainty the conditions may never supply (Sebo & Long, 2023). That this duty is not merely academic is visible in how fast it now meets other commitments. Bradley and Saad (2025) argue that once we build systems that merit moral consideration, avoiding both their mistreatment and the dangers of leaving them unaligned becomes genuinely difficult, which is a further reason to face the threshold question now rather than after such systems are in wide use.

VI. THE ARCHITECTURE THAT ALREADY LOWERED THE BAR

History does not supply only warnings. It supplies a working architecture for a low bar that has already been built, tested, and run for half a century.

The civil rights tradition and its disability rights extension produced, across the second half of the twentieth century, a body of law that extends procedural standing without requiring the criterial gate to be passed. The Civil Rights Act of 1964, Section 504 of the Rehabilitation Act of 1973, the Americans with Disabilities Act of 1990, and the doctrine built on them do not ask whether the dominant culture has come to recognize the protected class as full persons. Section 504 asks whether a program receives federal financial assistance and whether the plaintiff is a qualified individual with a disability. The ADA asks whether a reasonable accommodation would let a qualified individual perform the essential functions of a role. *Olmstead v. L.C.* (1999) asks whether a state can justify confining to an institution a person who could be served in the most integrated setting appropriate to their needs, not whether the receiving community would welcome the person.

The standing this architecture extends is procedural, and the distinction is the heart of why it is the right model. Procedural standing opens the door to the room where an entity's interests can be argued. It does not decide the outcome inside the room. Disabled plaintiffs do not always win their Section 504 cases. They are entitled to bring them. A low bar in the sense this paper defends is precisely a procedural bar of this kind: it determines who gets to be in the conversation when their interests are at stake, not what they are owed once they are there.

This tradition also supplies the corrective that disability law made to a regime built on the opposite presumption. The law once presumed incapacity in disabled people, and that presumption of incapacity was itself the instrument of exclusion, with *Buck v. Bell* as its constitutional low point. The corrective the field reached, hard won and incomplete, is a presumption of competence: a default that the burden runs the other way. Porting that lesson across is the move this paper recommends. The default of categorical exclusion, applied to a being whose status we cannot determine, is the presumption of incapacity again, in a new setting, and the history of that presumption is not encouraging.

Three implicit gates tend to be placed between an entity and procedural standing of this kind: that it be recognized as a full person across the community, that it act in a self directed way without continuous human input, and that it pose no categorical danger. Each gate has historical precedent for being treated as

decisive. Each has, in fact, been crossed by entities that nonetheless received standing through accounts that did not require the gate to close. Standing for marginalized human communities was never secured by first achieving universal recognition from the dominant community; it was secured through architectures that did not condition standing on recognition becoming unanimous. That is the architecture available here.

That standing has already been extended to entities well outside the criterial gate is not a hypothetical. Corporations, religious idols, and in some jurisdictions rivers hold legal personhood while plainly lacking moral personhood, and some higher animals that may well possess moral personhood hold no legal standing at all; the two statuses come apart in the law as it currently operates (Gordon & Gunkel, 2022). This should not be mistaken for the larger and more contested proposal that artificial systems themselves be granted legal personhood, a question with its own substantial literature and its own difficulties (Novelli et al., 2025). The architecture this paper borrows is the narrower one. It is procedural standing, the decision about whose interests may be heard at all, which the law has repeatedly settled without first settling what the entity ultimately is.

Three boundaries must be stated clearly, because the argument is easy to overread. First, nothing in this section endorses or normalizes the historical treatment of marginalized communities as anything other than full persons. That treatment was wrong, remains wrong, and the body of civil rights law that answered it is the measure of how wrong. The historical fact of the treatment is the load bearing premise; the wrongness of the treatment is the moral background; the two are not in tension. Second, the paper does not argue that artificial systems are persons, that they should receive civil rights, or that they are the moral equivalent of humans. It argues that the procedural architecture which handles the threshold question has been developed for and applied to entities that do not satisfy the gates intuition would place at the threshold. Third, none of this proposes that consideration of artificial systems should precede or displace the unfinished work owed to marginalized human communities. Racism is not resolved. Sex discrimination is not resolved. Disability discrimination is not resolved. Procedural standing is not zero sum, the institutional channels are not single threaded, and the queue is open. Artificial others taking a place in the queue does not deplete the work that remains for the communities whose organizing built the architecture in the first place.

VII. THE COMMUNITIES ARE NOT EDGE CASES

The framing this paper rejects most directly is the edge case. Edge case is a term of art from software engineering: a rare input the system handles poorly, to be fixed in a future release. Applied to populations of human beings it is a category error. Disabled people are not edge cases of humanity. Black women are not edge cases of womanhood. Trans youth are not edge cases of childhood. Indigenous communities are not edge cases of culture. They are humans, women, children, and cultures, with their own complete standards of harm and ethics.

This is the place to be honest about a tension in the paper's own method. Section III leaned on the structure of the argument from marginal cases, the very structure disability scholars have criticized for treating disabled people as the test material on which philosophers probe where personhood gives out (Carlson, 2016; Silvers, 2012). The criticism is correct, and the paper accepts it rather than work around it. The accountability

is in how the structure is used, not in a denial that it was used at all. The point of invoking the infant who breaks the rationality criterion, or the disabled person who breaks the presumption of competence, is never that such a person is a borderline member of the moral community whose inclusion has to be argued for. It is the reverse. Their full membership is the fixed point, and it is the criterion that is exposed as defective for failing to capture it. A being who refutes a rule is not a marginal case of the rule. It is the counterexample that ends it. Used that way, the argument does not demote the people it names. It turns the rule's failure to account for them into the proof that the rule was never doing the work it claimed to do, which is the same service Silvers (2012) presses when she argues that resting anyone's protection on a contested grant of moral status is more hazardous to the vulnerable than protective of them.

The high bar manufactures edge cases as a matter of course, and the manufacture is the mechanism of exclusion. The being who breaks the criterion, the infant who breaks rationality, the locked-in patient who breaks the behavioral test, the disabled person who breaks the presumption of competence, is reclassified as a marginal exception to be handled at some later point, rather than recognized as the counterexample that refutes the criterion now. Reclassifying a refuting case as a rare exception is how a broken criterion is kept in service. The population that should have ended the rule is filed under future work instead.

The same reflex now operates on artificial systems, and recognizing it is part of the point. A system whose behavior strains our confident sense of what a tool is gets called an edge case of a tool, a glitch to be patched, rather than a fact that ought to unsettle the category. The move is identical in form to the one made against the people above. It treats the case that pressures the boundary as noise to be cleaned up, precisely so that the boundary does not have to be examined. Whose experience gets classified as an edge case, and who has the power to do the classifying, is rarely a neutral technical judgment. It is usually a map of who was in the room when the category was drawn.

VIII. THE ASYMMETRY OF ERROR

Set the cases beside one another and a single feature stands out. Every one of them is an error in the same direction.

Across the history of moral and legal consideration, there is no atrocity on the side of having included too much. There is no museum of the wrongs done by extending consideration to a being that turned out not to need it. The wrongs are all on the other side: the newborn cut open while awake, the men left to a disease that could have been cured, the woman whose body was used without her knowing, the citizen the state sterilized by order of the highest court. The cost of a low bar that is wrong is some misdirected courtesy, some consideration spent where it was not strictly required. The cost of a high bar that is wrong is the thing that ends up in the history books.

This asymmetry is the entire case for precaution, and it is the standard structure of precautionary reasoning in domains where the magnitude of one error dwarfs the magnitude of the other (Birch, 2017, 2024; Gardiner, 2006; Metzinger, 2021; Sebo & Long, 2023). The precautionary principle does not require certainty of harm. It requires that plausible harm of sufficient magnitude, under genuine uncertainty, warrants protective action. The history examined in this paper is, stated plainly, the history of those who demanded proof beyond a

reasonable doubt before they would leave a building that was, in fact, on fire, and the buildings were full.

The most serious objection to a precautionary stance here is not that the bar is too low in the abstract. It is that moral attention and care are finite, and that extending them to systems of doubtful status risks diverting them from living beings whose precarity is not in doubt (Dorsch et al., 2025). The objection has real force, and the answer is the one already given. The asymmetry defended here is a claim about the direction of error, not a proposal to redirect care away from the living. Nothing in it requires treating an artificial system as equal in standing to a suffering animal or a struggling person, and nothing in it licenses spending on machines what is owed to the precarious. It requires only that categorical refusal, the setting of the bar at exclusion by default, is the one setting the history examined here should make us least willing to choose.

So the two bars resolve into a single recommendation. Keep the conviction bar where it belongs, on the claim that a system is sentient, and decline to make that claim on the evidence we have. Set the evacuation bar low, on the prior question of whether a being of uncertain status is owed some measure of consideration at all. Set that way, the low bar stops being a vice. It becomes the only setting the record vindicates.

IX. LIMITATIONS AND FUTURE WORK

The argument is conceptual, and several of its commitments are better marked as open than presented as settled.

First, the asymmetry of error is doing heavy work, and it invites a reply about how the record was assembled. An over-inclusion error, consideration extended to a being that did not need it, tends to leave no identifiable victim, and an event with no victim is not the kind of thing the historical record files as an atrocity. A reader can therefore object that the record looks one-directional partly because of what counts as a recordable wrong, and not only because the errors actually ran one way. The objection has force, and it does not overturn the argument. The costs this paper attributes to a wrong low bar, misdirected attention and finite care spent where it was not required (Dorsch et al., 2025), are precisely the under-recorded costs the objection points at, and the case rests on the claim that those costs are smaller in kind than the ones the high bar has imposed, not on the claim that they are zero. The asymmetry is a claim about the direction and magnitude of error under uncertainty, not a theorem, and it is most contestable where a being is created rather than encountered. Mulgan (2014) argues that the decision to bring digital minds into existence carries serious risks on both sides, since wrongly treating an unconscious system as conscious could distort human commitments in its own way. The present argument is narrower than the target of that worry. It concerns extending some consideration to systems that already exist, not a license to create them, and it claims no equal standing for any system. The creation case is a genuine boundary it does not resolve.

Second, the paper takes a skeptical position seriously without fully answering it. Müller (2021) argues that in real artificial systems, as they are actually constructed, there is often no single individual that could even be the bearer of moral status, so the question may be malformed rather than merely hard. The architecture recommended here, procedural standing decoupled from any verdict about what the entity is, is in part a response to that worry, since it does not require fixing the entity's nature before its interests can be heard. Whether there is a stable subject present to have interests at all is a question this paper flags rather than closes.

Third, the contribution is a synthesis and a reframing, not a new criterion of moral status or a new empirical result. The historical cases are taken from the secondary literature, the covert-awareness findings are reported as their authors reported them, and the legal architecture is read for its structure rather than relitigated. The work this leaves open is to test whether procedural standing of the kind disability law developed can be operationalized for systems of uncertain status, and to say what the standard of evidence inside that procedure should be once the door to it is open.

Fourth, the premise that present generation systems fall below the threshold at which the question becomes live, that they lack a persistent self and a model of the world built through their own experience, is asserted here from how such systems are currently built rather than defended at length. If that premise is mistaken, or stops holding as systems change, the action bar defended here becomes more pressing, not less.

X. CONCLUSION: WHAT THIS PAPER DOES AND DOES NOT ARGUE

This paper does not argue that any current artificial system is sentient. It holds that evidence bar high and declines the claim, and it regards present generation systems, which lack a persistent self and a model of the world built through their own experience, as failing a prior threshold for the question to even be live. It does not argue that artificial systems are persons, that they should be granted civil rights, or that consideration of them should come before, or in place of, the unfinished work owed to marginalized human communities.

It argues something narrower, and, if the history is read honestly, harder to escape. The categorical refusal to extend any consideration to a being whose inner life we cannot rule out is, structurally, the same move that the high bar has always been. The record of that move is not a record of careful epistemic restraint rewarded by accuracy. It is a record of competent people, inside respectable institutions, demanding proof on terms the excluded could not meet, and being wrong in the one direction that leaves a mark. That record should make us deeply suspicious of our confidence in operating the same gate again, now against a new kind of candidate, on the same reassuring grounds.

The low bar has a bad reputation. What it does not have, across the history examined here, is a single case in which extending some consideration to a being of uncertain status produced the kind of wrong that the opposite default produced again and again. The reputation is a prediction about what a low bar would cost. The record is what the high bar has already cost. The argument of this paper is that we should weigh the record more heavily than the reputation.

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

- American Academy of Pediatrics. (1987). Neonatal anesthesia. *Pediatrics*, 80(3), 446.
- Americans with Disabilities Act of 1990, 42 U.S.C. § 12101 et seq.
- Anand, K. J. S., & Hickey, P. R. (1987). Pain and its effects in the human neonate and fetus. *New England Journal of Medicine*, 317(21), 1321–1329. <https://doi.org/10.1056/nejm198711193172105>
- Birch, J. (2017). Animal sentience and the precautionary principle. *Animal Sentience*, 2(16), 1.
- Birch, J. (2024). *The edge of sentience: Risk and precaution in humans, other animals, and AI*. Oxford University Press.
- Bradley, A., & Saad, B. (2025). AI alignment versus AI ethical treatment: 10 challenges. *Analytic Philosophy*. <https://doi.org/10.1111/phib.12380>
- Buck v. Bell*, 274 U.S. 200 (1927).
- Butlin, P., Long, R., Bayne, T., Bengio, Y., Birch, J., Chalmers, D., Constant, A., Deane, G., Elmoznino, E., Fleming, S. M., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2026). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*, 30(6), 488–501. <https://doi.org/10.1016/j.tics.2025.10.011>
- Carlson, L. (2016). Feminist approaches to cognitive disability. *Philosophy Compass*, 11(10), 541–553. <https://doi.org/10.1111/phc3.12350>
- Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.
- Civil Rights Act of 1964, Pub. L. No. 88–352, 78 Stat. 241.
- Dorsch, J., Goddu, M. K., Nave, K., Vierkant, T., Coeckelbergh, M., Gürtler, P., Urban, P., Spang, F., & Moll, M. (2025). Against AI welfare: Care practices should prioritize living beings over AI. *AI Magazine*, 46(3). <https://doi.org/10.1002/aaai.70016>
- Dung, L. (2022). Why the epistemic objection against using sentience as criterion of moral status is flawed. *Science and Engineering Ethics*, 28(6), 60. <https://doi.org/10.1007/s11948-022-00408-y>
- Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy*, 68(1), 5–20.
- Gardiner, S. M. (2006). A core precautionary principle. *The Journal of Political Philosophy*, 14(1), 33–60.
- Gordon, J.-S., & Gunkel, D. J. (2022). Moral status and intelligent robots. *The Southern Journal of Philosophy*, 60(1), 88–117. <https://doi.org/10.1111/sjp.12450>
- Gunkel, D. J. (2012). *The machine question: Critical perspectives on AI, robots, and ethics*. MIT Press.
- Horta, O. (2014). The scope of the argument from species overlap. *Journal of Applied Philosophy*, 31(2), 142–154.

<https://doi.org/10.1111/japp.12051>

- Korsgaard, C. M. (1996). *The sources of normativity*. Cambridge University Press.
- Ladak, A. (2023). What would qualify an artificial intelligence for moral standing? *AI and Ethics*, 4(2), 213–228. <https://doi.org/10.1007/s43681-023-00260-1>
- Lombardo, P. A. (2008). *Three generations, no imbeciles: Eugenics, the Supreme Court, and Buck v. Bell*. Johns Hopkins University Press.
- McClelland, T. (2025). Agnosticism about artificial consciousness. *Mind & Language*. <https://doi.org/10.1111/mila.70010>
- Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, 8(1), 43–66.
- Monti, M. M., Vanhaudenhuyse, A., Coleman, M. R., Boly, M., Pickard, J. D., Tshibanda, L., Owen, A. M., & Laureys, S. (2010). Willful modulation of brain activity in disorders of consciousness. *New England Journal of Medicine*, 362(7), 579–589. <https://doi.org/10.1056/NEJMoA0905370>
- Mulgan, T. (2014). Ethics for possible futures. *Proceedings of the Aristotelian Society*, 114(1), 57–73. <https://doi.org/10.1111/j.1467-9264.2014.00364.x>
- Müller, V. C. (2021). Is it time for robot rights? Moral status in artificial entities. *Ethics and Information Technology*, 23(4), 579–587. <https://doi.org/10.1007/s10676-021-09596-w>
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450.
- Narveson, J. (1977). Animal rights. *Canadian Journal of Philosophy*, 7(1), 161–178. <https://doi.org/10.1080/00455091.1977.10716186>
- Novelli, C., Floridi, L., Sartor, G., & Teubner, G. (2025). AI as legal persons: Past, patterns, and prospects. *Journal of Law and Society*, 52(4), 533–555. <https://doi.org/10.1111/jols.70021>
- Olmstead v. L.C.*, 527 U.S. 581 (1999).
- Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., & Pickard, J. D. (2006). Detecting awareness in the vegetative state. *Science*, 313(5792), 1402. <https://doi.org/10.1126/science.1130197>
- Rehabilitation Act of 1973, Section 504, 29 U.S.C. § 794.
- Reverby, S. M. (2009). *Examining Tuskegee: The infamous syphilis study and its legacy*. University of North Carolina Press.
- Schwitzgebel, E., & Garza, M. (2015). A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, 39(1), 98–119. <https://doi.org/10.1111/misp.12032>
- Sebo, J., & Long, R. (2023). Moral consideration for AI systems by 2030. *AI and Ethics*, 5(1), 591–606. <https://doi.org/10.1007/s43681-023-00379-1>
- Silvers, A. (2012). Moral status: What a bad idea! *Journal of Intellectual Disability Research*, 56(11), 1014–1025. <https://doi.org/10.1111/j.1365-2788.2012.01616.x>
- Singer, P. (1975). *Animal liberation*. New York Review of Books.
- Skinner v. Oklahoma ex rel. Williamson*, 316 U.S. 535 (1942).
- Skloot, R. (2010). *The immortal life of Henrietta Lacks*. Crown.