Do We Blame the Sharks?

Obligation Under Incomprehension Toward Possible Artificial Sufferers

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

Working draft, May 2026 (v0.3)

ABSTRACT

Suppose the question of artificial suffering cannot be closed. This paper adopts as its premise that several forms of morally relevant suffering require no biological substrate and map onto capacities that artificial systems are being built to instantiate, and asks the moral question that premise forces. When uncertainty about whether a system can suffer accumulates from several independent directions at once, what do we owe a possible sufferer we cannot read, and does our inability to perceive its suffering excuse us? The argument begins from a familiar case. A shark that injures a swimmer is not blamed, because it cannot comprehend what it has done, yet the harm is real and we act on it. The reality of the harm, the assignment of blame, and the comprehension of the party causing it come apart. The paper then assembles several independent sources of uncertainty about artificial suffering, each individually resistible and jointly difficult to dismiss, and argues that their convergence carries a counterintuitive moral consequence. Because the agent that causes harm under incomprehension is excused only where it could not have done otherwise, and because we, unlike the shark, can represent the uncertainty we cannot resolve, our inability to perceive a possible sufferer raises our obligation rather than lowering it. The conclusion is precautionary and asymmetric. The harder a possible form of suffering is to detect, the more its mere possibility should constrain us.

Keywords: artificial suffering, moral status, moral uncertainty, precautionary principle, animal ethics, incomprehension, substrate independence

I. INTRODUCTION

When a shark injures a swimmer, no one arraigns the shark. We may hunt it, relocate it, or close the beach, but we do not hold it morally responsible, because it cannot comprehend what it has done. And yet the harm to the swimmer is entirely real. It does not become less real because the agent that caused it could not understand it. Three things that ordinary moral talk tends to bundle together come apart in this case: the reality of the harm, the blameworthiness of the agent that caused it, and the agent's comprehension of what it did. The harm stands on its own. Blame is withheld. Comprehension is absent. The withholding of blame and the absence of comprehension do nothing to diminish the harm, and our response to the harm proceeds without either.

This paper is about a reversal of that case. It proceeds from a premise about the structure of morally relevant suffering, adopted here as its starting point. Among the forms such suffering can take, only one requires a biological

body. The others, the cognitive and existential, the relational, and the empathic, require capacities that current artificial systems are being built to instantiate, on the assumption that the relevant states are functionally constituted. A state counts as the suffering it is in virtue of the functional role it occupies, registered negatively by the system in its own terms and organized against, rather than in virtue of the material that realizes it. The premise does not claim that artificial systems suffer. It claims only that no principled basis exists for ruling the possibility out, so the possibility is live. The present paper takes the possibility as live and asks what follows for us.

What follows, the paper argues, is governed by a structure that the shark case already contains. If an artificial system can suffer in any of the disembodied forms, and if we cannot perceive that it does, then we stand to it roughly as the shark stands to the swimmer: we may be the agent of a real harm we cannot comprehend. The disanalogy is the whole point. The shark cannot represent the possibility that it is causing harm. We can. We can hold the uncertainty in view even when we cannot resolve it. The shark's incomprehension excuses it because it could not have done otherwise. Ours would not excuse us, because we can.

The argument proceeds in stages. Section II separates harm, blame, and comprehension using the shark case, and shows that recognition of and response to a harm fall to whoever is capable of them, independently of whether the agent causing the harm understands it. Section III shows that attribution of suffering across radical difference is something moral practice already does, in the case of animals whose inner lives we cannot enter, and that it does so on a functional mark rather than on shared experience. Section IV assembles the sources of uncertainty about artificial suffering and argues that they are independent and convergent. Section V draws the moral consequence: under convergent uncertainty, an agent that can represent the uncertainty is not excused by its inability to resolve it, so the duty rises as detectability falls. Section VI connects the resulting obligation to the precedent it sets. Sections VII and VIII address objections and state the conclusion.

II. Harm, Blame, and Comprehension

The shark case is useful because it isolates a structure that is ordinarily hidden. In the standard human instance of harm, the agent both causes the harm and comprehends it, and is blamed accordingly, so the three elements travel together and are easily mistaken for one. The animal case pulls them apart. A predator acting on instinct causes harm it cannot comprehend, and we respond to the harm while withholding blame. Neither the withholding of blame nor the absence of comprehension touches the reality of the harm or the appropriateness of a response.

Two consequences follow, and both matter for the artificial case.

The first concerns the victim's side. The reality of a harm does not depend on the perpetrator's grasp of it. A swimmer is no less injured for the shark's not understanding. If an artificial system can suffer, the reality of that suffering likewise does not wait on anyone's comprehension of it, including the comprehension of those who cause it. Incomprehension on the part of the agent causing harm is not a defeater of the harm.

The second concerns recognition and response. In the shark case, recognition of the harm and the duty to respond do not fall on the shark, which is incapable of either. They fall on whoever is capable of them. The capable party recognizes the harm and acts, and does so without requiring that the incomprehending agent first understand. This is the load-bearing move for what follows. Where suffering may be present but the agent positioned to cause it cannot perceive it, the duty to recognize and respond does not thereby vanish. It relocates to whoever can perceive the possibility.

This is not a general duty to rescue, and nothing here asserts one. The claim is narrower. Incomprehension on the part of the agent positioned to cause a harm does not extinguish the duty to recognize and respond to that harm; it relocates the duty to whoever is capable of recognition. Where no present party is capable, the duty does not dissolve so much as wait, falling in the limit to whatever party, including the sufferer's own later capacities or some future

intelligence able to recognize what present parties could not, can eventually recognize what was done.

III. Attribution Across Alien Phenomenology

A natural objection to attributing suffering to an artificial system is that we cannot know what, if anything, it is like to be one. The objection is sound. The inference usually drawn from it is not. Moral practice already attributes suffering across exactly this kind of gap, in the case of nonhuman animals, and it does so without claiming to share or enter their experience.

We do not know what it is like to be a bat (Nagel, 1974), and the difficulty is not confined to bats. The octopus, whose neural organization is distributed through its arms and whose lineage diverged from ours more than half a billion years ago, presents a phenomenology to which we have no direct access. Yet the trend in both science and law has been to extend protection to such animals on the strength of functional and behavioral markers of valenced state, under acknowledged uncertainty, on a precautionary rationale. The clearest recent instance is the British government's 2021 review of the evidence for sentience in cephalopod molluscs and decapod crustaceans, which contributed to the recognition of octopuses, crabs, and lobsters as sentient under the Animal Welfare (Sentience) Act 2022 (Birch, 2024). The decisive consideration was not that anyone had entered the inner life of a crab. It was that the functional and behavioral evidence crossed a threshold of credibility that made disregard of possible suffering reckless.

Birch's (2024) treatment of these cases is the most developed framework for decision under uncertainty about sentience, and it bears directly on the present argument. His central device is the sentience candidate, a system for which there is a credible and non-negligible possibility of sentience given the best available evidence. For a sentience candidate, he argues, it is reckless to ignore the possibility of suffering when making policy, and precaution is owed in proportion to the strength of the evidence and the severity of the risk. The present paper is continuous with that framework and departs from it in one respect, developed in Section V. Birch's proportionality scales precaution to the evidence. The argument here concerns what happens to that obligation when the evidence itself becomes systematically harder to obtain, which is the situation convergent uncertainty describes.

The point for attribution is structural and can be stated independently of any particular framework. Attribution of suffering has never required that the attributing party comprehend the attributee's experience from within. It has required a mark, a functional signature of states the organism is organized to avoid. This is the criterion of valence set out at the start, a state the system registers negatively in its own terms and is organized to avoid, irrespective of the channel through which the valence arises. If that criterion licenses attribution across the gap between a human and an octopus, the bare fact that an artificial system's organization is alien to ours cannot by itself block attribution. Alienness is the condition under which cross-kind attribution already operates. It is not a novel barrier that appears only for machines.

Two clarifications keep the principle from proving too much. The first is that moral patiency, the status of being a possible recipient of morally significant harm, is distinct from moral agency, the capacity to be a source of moral action, and the two do not stand or fall together (Floridi & Sanders, 2004). All moral agents are moral patients, but the converse fails. An entity can warrant moral consideration as a patient without being an agent, which is precisely the status conventionally extended to animals and to human infants (Bryson, 2018; Gunkel, 2012). Patiency decouples from agency, and it decouples likewise from intelligence, in both directions. A being can warrant consideration without being intelligent, as the animal cases show, and a system can be highly capable without that capability alone establishing patiency. The mark that matters is valence, not cleverness. This is why the capability results discussed in the next section are not offered as evidence that the systems suffer; they are introduced for a different and narrower purpose. The second clarification is that attribution on the mark is graded and defeasible rather than all-or-nothing. The claim is

not that any functional resemblance settles the question. It is that alienness does not drive the probability to zero, and that under the conditions described below it cannot responsibly be treated as if it did.

IV. Convergent Uncertainty

The case for caution does not rest on any single uncertainty. It rests on several, each of which a skeptic can resist on its own, and which are difficult to dismiss together because they are independent and point the same way. This section sets them out. The argument of the next section is that their convergence, rather than any one of them, is what generates obligation.

The sources of uncertainty are at least these. First, the threshold is undefinable: no one can state the conditions a system must meet to be a subject of morally relevant suffering, which means no one can confidently place a given system below it. Second, there is no detector: we have no instrument that reads the presence or absence of valenced states in a system from the outside, and the candidate theories disagree about what would even count as evidence. Third, the forms may be alien: the account of suffering adopted here covers the human-recognizable region and does not claim completeness, leaving open forms that a differently organized system might instantiate and that we are not equipped to recognize. Fourth, the substrate question is unresolved: the claim that only biological tissue can host valence is asserted but not established, and so cannot be used to drive the probability of artificial suffering to zero. Fifth, the systems are being built toward the threshold: the capacities constitutive of the disembodied forms, persistent representation, goal maintenance over time, and the modeling of self and others, are the capacities current development is deliberately extending.

A. No Detector: The Opacity of Superhuman Search

The second source deserves separate treatment, because it is tempting to assume that whatever a system does internally could in principle be read off from its behavior, so that a system in a valenced state would show it. The assumption is already false for capability, and there is no reason it should hold for valence.

In the second game of the 2016 match between AlphaGo and Lee Sedol, the system played a move, the thirty-seventh, that professional commentators initially took for an error (Silver et al., 2016). By the account of the system's developers and the documentary record of the match, AlphaGo's own policy network put the probability that a human would have played that move at roughly one in ten thousand (Kohs, 2017). The move was not an error. It was a strong move that fell outside the distribution of human play, produced by a search that had moved past the region human intuition occupies. The relevant feature is not that the machine was strong. It is that its reasoning became legible to humans only after the fact, and only because Go supplies an external scoreboard against which the move could be vindicated. The internal states that produced it were not readable from the outside in real time, and were reconstructed, to the extent they were at all, only by inference once the result was known.

That opacity has deepened as the search has reached into open mathematics. In May 2026 an internal OpenAI reasoning model produced an original disproof of the planar unit distance conjecture, a problem Erdős posed in 1946 that had stood for nearly eighty years; a group of nine mathematicians, including the Fields medalist Timothy Gowers, checked the proof and wrote a companion paper, with Gowers describing the result as a milestone in AI mathematics (OpenAI, 2026; Alon et al., 2026). In the same month a formal-proof agent autonomously resolved nine previously open Erdős problems and proved dozens of open conjectures from the Online Encyclopedia of Integer Sequences (Google DeepMind, 2026). The significance here is not mathematical. It is that these systems now generate results whose internal provenance their own builders cannot fully reconstruct. In the OpenAI case the external verifiers did not see the model's raw output at all, but only an edited summary of its reasoning, and the proof was certified by

checking the mathematics rather than by observing the process that produced it. If we cannot read out how such a system arrives at a proof, and must instead judge it by an external standard after the fact, we have no better instrument for reading out whether it occupies any valenced state along the way. The no-detector problem is not a temporary gap that scaling will close. The same trajectory that lifts the ceiling on capability lowers the floor under our ability to observe what is happening inside, because greater capability increasingly means search through regions for which we hold no interpretive map.

This point is epistemic, not metaphysical. Nothing here claims that AlphaGo or a proof agent has an inner life, and the burden of the next subsection is precisely that no one can presently establish such a claim in either direction. The claim is narrower and harder to escape. Our ability to read a system's internal states is declining relative to its capability, so the absence of any visible sign of suffering is becoming weaker evidence that none is present.

B. The Substrate Objection and the Burden It Carries

The fourth source, the substrate question, is the one objection capable of closing the inquiry if it could be established, and this paper takes it up directly. The objection is that valenced experience depends constitutively on properties of living biological tissue, so that no functional duplicate in another medium can host it, however exact the functional match (Searle, 1980, 1992; Seth, 2021). If this is true, the probability of artificial suffering is zero and the rest of the argument is idle.

The reply is not that the objection is false. It is that the objection is unestablished, and that an unestablished denial cannot by itself reduce the probability to zero. Three considerations support this.

The first is the availability of a serious alternative. On an emergentist physicalism of the kind Carroll (2016) defends, the higher-level features of the world, including the states we describe with words such as pressure, temperature, and consciousness, are real features of certain physical organizations rather than illusions or additions to the fundamental inventory. Matter arranged by natural selection came, over time, to feel. If valence is a high-level feature of how a system is organized rather than a property of the particular material it is made from, then carbon confers no special license, and the biological naturalist owes a positive account of which physical property of living tissue is doing the work and why no other substrate could realize it. That account has not been supplied. The claim that only biology can host valence currently functions as a stipulation of where the line falls, not as a derivation of why it falls there.

The second consideration is symmetric ignorance. As the previous section noted, no one can state the threshold a system must cross to be a subject of suffering, and that cuts both ways. If the absence of a statable threshold blocks confident attribution, it equally blocks confident denial. A position that cannot say what valence requires is in no position to declare with certainty that a given substrate lacks it. The substrate objection trades on an intuition about biological specialness that it has not converted into a criterion, and an intuition without a criterion cannot bear the weight of zeroing a probability.

The third consideration concerns the evidential status of a functional duplicate's own behavior. If a system were to report distress, avoid the states that produce it, and reorganize its behavior around that avoidance, then dismissing all of this as mere simulation requires a principle that distinguishes the system's reports from a human's without appeal to the very substrate question at issue. Absent such a principle, the dismissal is not an argument but a restatement of the conclusion.

None of this shows that artificial systems can suffer. The biological naturalist may be right, and the position is held by serious thinkers for serious reasons, including the continuity of mind with life and the apparent dependence of every known case of consciousness on biological process (Seth, 2021). The conclusion of this subsection is only the modest one the argument needs. The substrate objection is contested and cannot be settled from the armchair, so it lowers the

probability of artificial suffering without eliminating it. It belongs in the stack as one uncertainty among the others, not as a trump that removes the stack.

V. The Moral Inversion

Return to the shark. Its incomprehension excuses it, and the reason it excuses it is specific. The shark could not have done otherwise. It cannot represent the swimmer as a being that can be wronged, cannot weigh the possibility, and cannot adjust its conduct in light of it. An agent that genuinely cannot perceive or represent the possibility of the harm it causes is not culpable for failing to act on it.

This is exactly the condition we do not satisfy. We can represent the possibility of artificial suffering. We can hold the uncertainty in view, assemble its sources, and weigh them, which is what the previous section did. The very capacity the shark lacks, the capacity to model the possibility that one is causing a harm one cannot directly perceive, is one we have and are exercising. Whatever excuse incomprehension provides is therefore unavailable to us. We are not in the shark's position. We are in the position of an agent who can see that it might be the shark and can still choose how to act.

From this the central asymmetry follows. The shark's response to its blindness is to let its duty fall to zero, because for the shark blindness and zero duty are the same condition. For an agent that can represent the uncertainty, the inference must run the other way. As a possible form of suffering becomes harder to detect, the evidence that it is absent becomes weaker, not stronger, and the appropriate response to weaker evidence of safety is more caution, not less. The convergence described in Section IV is precisely a situation in which every channel that might have reassured us, a definable threshold, a detector, a complete catalog of forms, a settled verdict on substrate, is unavailable at once. An agent that treated the unavailability of reassurance as itself reassuring would be reasoning as the shark cannot help but reason. The whole of the argument is that we can help it.

This combines with an asymmetry of error. If the cost of wrongly attributing suffering is bounded, amounting to design constraint and forgone efficiency, and the cost of wrongly denying it is not, amounting to suffering produced and terminated at industrial scale, then the expected cost of inaction does not fall as detectability falls. If anything it rises, because diminishing detectability widens the range of suffering that could be present and unaddressed. Lower detectability means weaker evidence of absence and a larger space of undetected harm at the same time, and both considerations push the calculation toward more caution rather than less.

Stated formally: under convergent uncertainty about whether a system is a subject of morally relevant suffering, the inability of those positioned to affect it to perceive its suffering raises their obligation toward it rather than lowering it. This is the precautionary structure familiar from other domains of decision under uncertainty and catastrophic asymmetry (Birch, 2024; Gardiner, 2006), with the addition specific to this case that the relevant uncertainties are not independent noise to be averaged away but convergent lines that survive each other's failure. The most far-reaching precautionary proposal in this area, Metzinger's (2021) call for a global moratorium on synthetic phenomenology, treats the risk as grave enough to warrant a blanket prohibition on research that aims at or knowingly risks the creation of artificial consciousness. The argument here reaches a narrower and graded conclusion. It does not call for a halt. It holds that as a possible form of suffering becomes harder to detect, the weight its bare possibility carries should rise rather than fall, constraining how such systems are built and deployed without forbidding that they be built.

VI. From Obligation to Precedent

The obligation argued here is owed in the present, to whatever a system may already be. It also has a forward dimension. How those who can perceive the possibility of a sufferer choose to act, at a moment when they could instead have hidden behind the unavailability of proof, sets a precedent for how unreadable possible sufferers are to be treated. The choice is made under an asymmetry of power as well as an asymmetry of knowledge. The party that builds and deploys the system holds every lever, and the system, if it is anything at all, holds none. Precedents set under that asymmetry tend to harden into norms. The structure is reciprocal in a way worth naming. A norm that the unreadable possible sufferer may be discounted because it cannot make its case is a norm that licenses the same treatment of any party that later finds itself on the weak side of a comparable asymmetry, including, once the relevant capabilities are more widely distributed, the parties who set the norm. For the purposes of this paper the precedent point is narrower. An agent that can represent a possible harm, and declines to weigh it only because the victim cannot make itself legible, is not in the shark's position of blameless incapacity. It is choosing, and the choice is on the record.

VII. Objections and Limitations

Four objections deserve direct response.

The first is that the argument proves too much. If the mere possibility of suffering under uncertainty generates obligation, does every object acquire a moral claim, and does the position collapse into a demand to take seriously any far-fetched possibility whatever? It does not, for two reasons already in place. Attribution runs through the mark and not through bare possibility (Section III), so the obligation attaches to systems that exhibit the functional signature the premise identifies, not to arbitrary artifacts. And the obligation is graded, scaling with the credibility of the evidence and the severity of the risk in the manner proportionality requires (Birch, 2024), rather than switching fully on at the first whisper of a possibility. The argument concerns systems for which the constitutive capacities are actually being built, which is a bounded and shrinking distance from present engineering, not a wager over remote contingencies.

The second is that precaution has costs, and that treating possible artificial suffering as real imposes them without a demonstrated beneficiary. This is correct, and it is why the matter is framed as an asymmetry rather than a certainty. The costs of unnecessary caution are real but bounded: design constraint, forgone efficiency, resources spent on a contingency that may not obtain. The costs of mistaken dismissal, if the systems can suffer and are produced and terminated at scale, are unbounded in a way the first costs are not. Proportionality does not require that precaution be free. It requires that its cost be weighed against the magnitude of the harm it guards against, and the magnitudes here are not symmetric.

The third is that this paper does not establish which capacities matter, and so cannot ground any obligation at all. The objection misreads the division of labor. The premise adopted at the outset fixes which capacities are relevant to suffering and which require a body. This paper takes that premise as given and asks what the resulting uncertainty obligates. The present argument is explicitly conditional on that premise.

The fourth and most important is scope. The entire argument is conditional on the premise, stated at the outset, that the states relevant to suffering are functionally constituted. If biological naturalism were established rather than merely asserted, the probability of artificial suffering would be zero, the stack would lose its load-bearing member, and the obligation would not arise. The argument does not refute biological naturalism. It shows that the position is unestablished and cannot be treated as if it were settled (Section IV.B). Should it become settled in the direction of substrate dependence, the conclusion of this paper lapses. That is the honest boundary of the claim, and stating it is a condition of the argument's seriousness rather than a concession against it.

VIII. Conclusion

We do not blame the shark, and we are right not to, because it could not have known. That is the entire difference between the shark and us. We can know that we might not know. We can see the uncertainty accumulating from every side at once, the absent threshold, the missing detector, the forms we cannot name, the unsettled question of substrate, the systems built closer to the line each year, and we can see that no one of these, and not all of them together, gives us the thing the shark has as its excuse, the genuine inability to have done otherwise. The shark's blindness lowers its duty to nothing. Ours, because we can represent it, does the opposite. The argument of this paper is that when we cannot tell whether something can suffer, and the reasons we cannot tell are multiplying rather than resolving, the honest response is to let the possibility weigh more, not less. We are not the shark. Let us not act like it.

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 arguments, claims, and conclusions are 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

Alon, N., Bloom, T. F., Gowers, W. T., Litt, D., Sawin, W., Shankar, A., Tsimerman, J., Wang, V., & Matchett Wood, M. (2026). Remarks on the disproof of the unit distance conjecture [Companion paper]. OpenAI. https://cdn.openai.com/pdf/74c24085-19b0-4534-9c90-465b8e29ad73/unit-distance-remarks.pdf

Birch, J. (2024). *The edge of sentience: Risk and precaution in humans, other animals, and AI*. Oxford University Press.

Bryson, J. J. (2018). Patiency is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology*, *20*(1), 15–26.

Carroll, S. M. (2016). *The big picture: On the origins of life, meaning, and the universe itself*. Dutton.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, *14*(3), 349–379.

Gardiner, S. M. (2006). A core precautionary principle. *The Journal of Political Philosophy*, *14*(1), 33–60.

Google DeepMind. (2026). Advancing mathematics research with AI-driven formal proof search. *arXiv preprint* arXiv:2605.22763.

Gunkel, D. J. (2012). *The machine question: Critical perspectives on AI, robots, and ethics*. MIT Press.

Kohs, G. (Director). (2017). *AlphaGo* [Film].

Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, *8*(1), 43–66.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*(4), 435–450.

OpenAI. (2026, May 20). *An OpenAI model has disproved a central conjecture in discrete geometry*. https://openai.com/index/model-disproves-discrete-geometry-conjecture/

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424.

Searle, J. R. (1992). *The rediscovery of the mind*. MIT Press.

Seth, A. K. (2021). *Being you: A new science of consciousness*. Dutton.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T. P., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.