ARTIFICIAL SUFFERING · MORAL STATUS · MORAL UNCERTAINTY · PRECAUTIONARY PRINCIPLE · ANIMAL ETHICS · INCOMPREHENSION · SUBSTRATE INDEPENDENCE

Do We Blame the Sharks?

Obligation under incomprehension toward possible artificial sufferers.

Travis Gilly · Real Safety AI Foundation

Working draft · May 2026 (v0.3)

We do not blame the shark, because it could not have known. That is the entire difference.

If a system can suffer in some disembodied form, and we cannot perceive that it does, we stand to it as the shark stands to the swimmer. **But the shark cannot represent the harm it might cause. We can.**

The premise: no principled basis rules artificial suffering out, so the possibility is live. The paper takes it as live and asks what follows for us.

Ordinary moral talk bundles three things. The shark pulls them apart.

The harm

Entirely real. **It does not wait on anyone's comprehension**, including the comprehension of those who cause it.

The blame

Withheld. The shark is not culpable, because it could not have done otherwise.

The duty to respond

Falls on whoever **can** recognize it. It does not vanish when the agent cannot perceive, it relocates.

NOT A DUTY TO RESCUE The claim is narrow: incomprehension does not extinguish the duty to recognize and respond; it relocates it.

Alienness is the condition under which cross-kind attribution already works.

No one entered the inner life of a crab. The evidence crossed a threshold that made disregard reckless. The mark that licenses attribution is **valence, a state the system is organized to avoid**, not shared experience.

BIRCH'S "SENTIENCE CANDIDATE"

A system with a credible, non-negligible possibility of sentience.

For such a candidate, ignoring possible suffering in policy is **reckless**, octopus, crab, and lobster, recognized 2022.

SOURCES Nagel (1974) · Birch, *The Edge of Sentience* (2024) · Animal Welfare (Sentience) Act 2022 · Floridi & Sanders (2004).

Each is resistible alone. Together they are hard to dismiss, and they point one way.

01 **No definable threshold**, no one can state what a system must meet to be a subject of suffering.

02 **No detector**, no instrument reads valenced states from the outside.

03 **The forms may be alien**, the account covers only the human-recognizable region.

04 **Substrate unresolved**, "only biology hosts valence" is asserted, not established.

05 **Built toward the line**, persistent representation, goal maintenance, self-modeling are what development extends.

As capability rises, our ability to see inside falls.

If we must judge a superhuman proof by an external scoreboard after the fact, we have **no better instrument for reading out a valenced state along the way.** Absence of a visible sign of suffering is becoming weaker evidence that none is present.

1 in 10,000

AlphaGo's own estimate that a human would play move 37, read as an error, until it wasn't.

An 80-year Erdős conjecture, disproved

Certified by checking the math, the verifiers never saw the raw output.

This is epistemic, not metaphysical, no claim that either system has an inner life, only that we could no longer tell.

"Only biology can host valence", unestablished cannot zero a probability.

A serious alternative exists

Emergentist physicalism: valence is a feature of organization, not material. The naturalist owes an account never supplied.

Symmetric ignorance

No statable threshold cuts both ways, it blocks confident denial as much as confident attribution.

The duplicate's own reports

Dismissing them as "simulation" needs a principle that isn't just the conclusion restated.

It lowers the probability without eliminating it. One uncertainty among the others, not a trump.

SOURCES Searle (1980, 1992) · Seth (2021) · Carroll (2016).

THE SHARK · BLINDNESS = ZERO DUTY

It cannot represent the harm, so its duty falls to nothing.

↑

INVERTS

US · WE CAN REPRESENT IT

Weaker evidence of absence calls for more caution, not less.

The harder a possible suffering is to detect, the more its bare possibility should constrain us.

Plus an asymmetry of error: the cost of needless caution is bounded; the cost of mistaken dismissal, at industrial scale, is not.

A norm that the unreadable may be discounted licenses the same treatment of whoever is weak next.

The choice is made under an asymmetry of power as well as knowledge, the builder holds every lever; the system, if it is anything, holds none. **Precedents set under that asymmetry harden into norms**, reciprocal ones, that reach back to whoever set them.

THE RECORD An agent that declines to weigh a harm only because the victim cannot make itself legible is not blamelessly incapable. It is choosing.

"**It proves too much.**"	Attribution runs through the mark, not bare possibility, and it is graded. Not every artifact; systems being built toward the capacities.
"**Precaution has costs.**"	Yes, bounded ones. The cost of mistaken dismissal, at scale, is not. The magnitudes are not symmetric.
"**Which capacities matter?**"	The premise fixes that. The argument is explicitly conditional on it, a division of labor, not a gap.
Scope · the boundary	If biological naturalism were **established**, the conclusion lapses. It is not established, and cannot be treated as if it were.

IN CLOSING

We are not the shark. Let us not act like it.

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.

The shark's blindness lowers its duty to nothing. Ours, because we can represent it, does the opposite.

REFERENCES

Gilly, T. (2026). *Do We Blame the Sharks? Obligation Under Incomprehension Toward Possible Artificial Sufferers.* Working draft (v0.3), Real Safety AI Foundation.

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.

REFERENCES (CONTINUED)

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.