REAL SAFETY AI FOUNDATION · WORKING PAPER

MACHINE CONSCIOUSNESS · MORAL STATUS · PATTERN MATCHING · ANIMAL COGNITION · PROBLEM OF OTHER MINDS · PRECAUTIONARY PRINCIPLE · WORLD MODELS · LARGE LANGUAGE MODELS

The Bill Comes Due at the Dog

Conditioning, pattern matching, and the limits of certainty about machine minds.

Travis Gilly · Real Safety AI Foundation

Working paper · June 2026 (v1.2)

IT IS JUST,

predicting the next token. Matching patterns in its training data. A stochastic parrot.

The sentence is offered as though it closed the question, and most of the time it is allowed to. This paper argues that it closes nothing, and the argument it relies on is not new. That is the point. It has been available in the philosophy of animal minds for a century.

SOURCES Bender, Gebru, McMillan-Major & Shmitchell (2021), *On the Dangers of Stochastic Parrots*, ACM FAccT, 610–623.

Almost nothing is *just* anything.

The move reduces a system to one of its true descriptions and treats the reduction as deflationary, the model is just a next-token predictor, just a function approximator, just autocomplete. But a thing can be accurately described at the level of its mechanism and remain, at another level, the very thing the mechanism was alleged to rule out. Naming the mechanism names a fact. It does not settle the question the fact was invoked to settle.

A brain
is just a large collection of cells exchanging ions.

A symphony
is just air pressure varying over time.

A person
is, among other things, just an arrangement of meat.

SOURCES Aaronson, "Justaism," as discussed in Shevlin (2026), *Contra Chiang on Machine Consciousness.*

PART ONE

Pavlov as Disproof

The oldest case in the experimental record settles the matter, against the dismissal.

I

Pavlov's dog salivates because a stimulus has been paired with a reward until the association is learned. That is associative pattern matching in its most literal form, and no one concludes the dog feels nothing.

The bell, bound by repetition to food. Salivation, on its own. Pattern matching, in the most literal sense the words can carry.

"And yet no one stands over the conditioned animal and concludes that there is nothing it is like to be the dog, that the animal is an empty mechanism with the lights off."

We grant the dog an inner life, and we grant it on grounds that have nothing whatever to do with the mechanism of its learning. The mechanism is simply not where the question of the inside is decided.

SOURCES Pavlov (1927), *Conditioned Reflexes.* · Nagel (1974), *What Is It Like to Be a Bat?*, Phil. Review 83(4).

The presence of associative pattern matching does not entail the absence of an inner life.

WHAT THE CLAIM IS NOT

Not that conditioning is consciousness, nor that associative learning constitutes or explains the felt life of the organism. That would be its own reductionism, and a false one.

WHY IT HOLDS

One uncontested case is all a negative claim requires. In the dog, the mechanism is unmistakably present and the inner life is granted. Mechanism and phenomenology are logically independent: naming the first cannot settle the second.

SOURCES Nagel (1974), *What Is It Like to Be a Bat?*, The Philosophical Review 83(4), 435–450.

Behaviorism, redeployed against silicon.

Behaviorism's ambition was to replace inner states with stimulus and learned response, dissolving the felt life of the organism into its observable contingencies. Almost no one now accepts that it succeeded; the reduction was rejected for animals and humans on principled grounds. The machine, we are now told, only does what it was trained to do, therefore there is no one home. It is the same inference, and it fails for the same reason.

Then, Watson & Skinner

"The inner life of a dog or a person is exhausted by its conditioning." Rejected, decisively, by Chomsky (1959).

↓

Now, against the machine

"Its outputs are shaped responses to inputs; therefore nothing is felt." The behaviorist inference, whatever else its user believes.

SOURCES Chomsky (1959), Review of *Verbal Behavior* by B. F. Skinner, Language 35(1), 26–58.

PART TWO

The Parity and the Bill

II

To stay certain, the skeptic owes a criterion the dog satisfies and the candidate lacks. The debt cannot be paid.

Anyone who asserts with certainty that there is nothing it is like to be a candidate system, while granting an inner life to the dog, has taken on a debt, and the body of this argument is an attempt to collect it.

Premise one	If associative pattern matching disqualifies a system from an inner life, then the dog, whose relevant learning is associative pattern matching, has no inner life.
Premise two	We do not accept that the dog has no inner life.
∴ Conclusion	**Associative pattern matching does not disqualify a system from having an inner life.**
Corollary	The operative result: that an artificial system is doing pattern matching cannot, by itself, establish that it has no inner life. The argument from the mechanism is unsound, because the same argument applied to a case we have already judged delivers a verdict we reject.

SOURCES Gilly, T. (2026), *The Bill Comes Due at the Dog*, §IV. Real Safety AI Foundation.

Each readmits the machine, or excludes the dog and the infant.

Substrate

ASSUMES THE ANSWER

A ground for presumption, not a demonstration that biology is necessary for feeling. To wield it as a disqualifier is to assume the very question at issue.

Behavior

CUTS THE OTHER WAY

By hypothesis the candidate's behavior is at least as rich as the dog's. One cannot grant the dog its inside on behavior and deny the machine in the face of behavior as rich.

Reasoning

EXCLUDES THE INFANT

The dog does not reason discursively either. If discursive reasoning were required, the dog fails it, and so does the human infant, and many beings whose inner lives no one doubts.

Language

EXCLUDES THE DOG

The dog has no language, and we grant it an inner life regardless. If language were the requirement, the dog is out and the prelinguistic infant is out.

This is the old argument from marginal cases, in a new setting: the features used to draw the boundary are possessed by some nonhumans and lacked by some humans.

SOURCES Sebo (2025), *The Moral Circle* (W. W. Norton); Sebo (2024), *Insects, AI Systems, and the Future of Legal Personhood.*

A lower credence is not a zero credence.

The probabilistic skeptic is right that the candidate warrants less credence than the dog. But a graded prior is still a probability, and a probability the evidence cannot drive to certainty is exactly the suspended judgment defended here. Reasoning honestly, the skeptic arrives at a candidate held inside the zone of reasonable disagreement, this paper's conclusion, stated in the vocabulary of credence. What they do not arrive at is certain absence.

AGNOSTICISM, THE WEAK CLAIM

"We do not know, and neither do you." Weak claims are hard to defeat.

CERTAIN ABSENCE, THE STRONG CLAIM

"There is nothing there at all." The strong claim is the one that owes the criterion, and the criterion does not exist.

SOURCES Birch (2024), *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI* (Oxford University Press).

PART THREE

Two Neighbors

One shares the premise and draws a different conclusion. The other shares the conclusion and needs this premise defended.

III

The relational turn in robot ethics, and recent precautionary approaches to moral status under uncertainty, the negative argument sits upstream of both.

Same premise, different altitude.

The relational turn shares the epistemic premise, inner states are opaque, and concludes that we should abandon properties and relocate moral standing in the relation itself. This argument keeps the inner question open. It is prior to that fork and narrower than either branch: it does not ground moral status, it forbids one shortcut to denying it. One can accept the conclusion here and remain a properties theorist, now appropriately agnostic, or go relational. The negative result sits upstream and constrains the choice from above.

WHY KEEP THE INNER QUESTION ALIVE

A relation can be warm and one-sided, directed at something with no inside at all. The sociological fact that we treat an entity as a fellow does not settle whether treating it badly harms anyone. For that, what the entity is continues to matter.

SOURCES Gunkel (2012, 2018, 2022); Coeckelbergh (2010); Coeckelbergh & Gunkel (2014); Gellers (2020).

Scale concern to the probability, not to a verdict that may never arrive.

Where the evidence of sentience is inconclusive, the precautionary principle gives the creature the benefit of the doubt and sets a deliberately low bar for protective measures. The same framework runs continuously from the common animal to the emerging artificial system.

THIS PAPER'S NARROW CONTRIBUTION

Sebo argues the uncertainty in general. This paper disarms the instrument by which the uncertainty is illegitimately dissolved.

People talk themselves out of caution by naming the mechanism, "it is only pattern matching", and treating the naming as resolution. The dog is the concrete demonstration that this resolution is no resolution at all.

SOURCES Birch (2017), *Animal Sentience and the Precautionary Principle.* · Long, Sebo, Butlin et al. (2024), *Taking AI Welfare Seriously* (arXiv:2411.00986). · Sebo (2025).

PART FOUR

The Boundary and the Candidate

IV

The argument is worthless if it sweeps in systems to which it does not apply. So what is, and is not, a candidate?

A current large language model is not a candidate for the question. The candidate is a different and approaching kind of system, and the step that produces it is the same step that first makes a real goal-directed agent possible.

The language model is not yet a someone at all.

In operation, a current model is a function from input to output, computed in a single forward pass. Beyond memory bolted on from outside, nothing of one exchange persists into the next as a state of the system itself. There is no standing subject that carries across time, and a system with no persistent self fails a threshold the argument requires before it can even begin. To insist on this is not to concede the dismissive case. It is to refuse the overclaim that would discredit the real one.

LESS THAN THE PAPERCLIP MAXIMIZER

The maximizer is at least a unified, persistent agent with a stable goal it pursues across time. It is someone, pointed in a catastrophic direction.

The language model lacks even that. The argument of this paper is not that the chatbot is alive.

SOURCES Bostrom (2014), *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press).

"A system that constructs and continually revises a model of the world through its own ongoing experience, in the way a human infant does."

THE BRIGHT LINE

Not a world model as such, a frozen system can hold a sophisticated one. The line is its experiential, continually revised construction through the system's own lived interaction over time. That is the developmental signature: a persisting subject learning from a world it is in.

A SCOPE CONDITION, NOT A DISQUALIFIER

It does not hold that a system which builds such a model thereby has an inside. It fixes the domain over which the parity reasoning runs, and stays exactly as agnostic within it. "Here is where the question applies", not "here the answer is no."

SOURCES Clark (2016), *Surfing Uncertainty.* · Birch (2024), on the sentience candidate.

The inside and the maximizer arrive on the same step.

BEFORE THE STEP

A frozen function

No persisting subject to be the locus of an inner life. No persisting agent to hold and pursue a goal against the world.

becomes

AFTER THE STEP, BOTH AT ONCE

A candidate for an inside

A persisting subject that learns from a world it is in. The moral question.

A real goal-directed agent

A persisting agent that pursues an end against the world. The safety question.

In the programs that model agency computationally, building a generative world model and the capacity for goal-directed action are two expressions of one architecture, not two. The capability that makes a system worth worrying about is bound to the capability that makes it a candidate for the inside.

SOURCES Basanisi, Brovelli & Cartoni (2020), PLOS Comp. Biol. 16(12), e1007579. · Matsumoto, Ohata & Tani (2023), Entropy 25(11), 1506.

"Unearned certainty is not free. It licenses a disposition whose costs fall due precisely at the moment we have handed the greater power to the party we were so sure was empty."

For some interval we will be the more capable party, the way a parent is more capable than the infant in its care. How we treat a candidate mind in that window is a precedent, what the powerful take themselves to owe the less powerful, laid down at the one time the lesson can still be set by us. Until someone names the criterion that keeps the dog in and the candidate out, the only honest stance is the one we already take toward the dog. We do not know what it is like to be it. And we do not pretend that we do.

SOURCES Gilly, T. (2026), *The Bill Comes Due at the Dog*, §VII. Real Safety AI Foundation.

Gilly, T. (2026). *The Bill Comes Due at the Dog: Conditioning, Pattern Matching, and the Limits of Certainty About Machine Minds.* Real Safety AI Foundation. Working paper (v1.2).

Basanisi, R., Brovelli, A., and Cartoni, E. (2020). A generative spiking neural-network model of goal-directed behaviour and one-step planning. *PLOS Computational Biology,* 16(12), e1007579.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery.

Birch, J. (2017). Animal sentience and the precautionary principle. *Animal Sentience,* 2(16).

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

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies.* Oxford University Press.

Chomsky, N. (1959). Review of *Verbal behavior* by B. F. Skinner. *Language,* 35(1), 26–58.

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* Oxford University Press.

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

Coeckelbergh, M., & Gunkel, D. J. (2014). Facing animals: A relational, other-oriented approach to moral standing. *Journal of Agricultural and Environmental Ethics,* 27(5), 715–733.

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

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

Gunkel, D. J. (2018). *Robot rights.* MIT Press.

Gunkel, D. J. (2022). The relational turn: Thinking robots otherwise. In J. Loh & W. Loh (Eds.), *Social robotics and the good life: The normative side of forming emotional bonds with robots.* Transcript Verlag.

Long, R., Sebo, J., Butlin, P., et al. (2024). *Taking AI welfare seriously* [Preprint]. arXiv. https://arxiv.org/abs/2411.00986

Matsumoto, T., Ohata, W., and Tani, J. (2023). Incremental learning of goal-directed actions in a dynamic environment by a robot using active inference. *Entropy,* 25(11), 1506.

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

Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex* (G. V. Anrep, Trans.). Oxford University Press.

Sebo, J. (2024). *Insects, AI systems, and the future of legal personhood* [Manuscript]. Retrieved from https://jeffsebo.net/

Sebo, J. (2025). *The moral circle: Who matters, what matters, and why.* W. W. Norton.

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

Shevlin, H. (2026). *Contra Chiang on machine consciousness* [Online essay]. Retrieved from https://www.polytropolis.com/p/contra-chiang-on-machine-consciousness