# The Bill Comes Due at the Dog:

# Conditioning, Pattern Matching, and the Limits of Certainty About Machine Minds

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

Working paper, June 2026 (v1.2)

**ABSTRACT**

A common move dismisses the possibility of an inner life in an artificial system by naming the mechanism that produces its behavior: the system is just predicting the next token, just matching patterns, just a stochastic parrot. This paper argues that the move settles nothing, and it does so through an argument that has been available in animal ethics for a century. Pavlovian conditioning is associative pattern matching, yet no one infers from the mechanism that a conditioned animal has no inner life. Mechanism and phenomenology are separate questions; naming the first does not answer the second. It follows that anyone who asserts with certainty that there is nothing it is like to be a candidate artificial system, while granting an inner life to the dog, owes a criterion that cleanly separates the two cases. The paper surveys the available criteria, namely substrate, behavior, reasoning, and language, and argues that each either readmits the systems the objector wants to exclude or excludes the animals and human infants we already include. The argument is strictly negative and strictly bounded. It does not claim that any system is conscious, and it explicitly excludes current large language models, which fail a prior threshold: a forward pass with no persistent self and no model of the world built through its own experience is not yet a candidate for the question. The candidate is the system that constructs and revises a world model through its own ongoing experience, in the way a human infant does. The paper locates this position against the relational turn in robot ethics, which uses the same epistemic premise to abandon the inner question, and against recent precautionary approaches to moral status under uncertainty, which supply the normative consequence the negative argument implies. It closes by arguing that certainty in the dismissive direction is not idle: it licenses a disposition toward candidate minds whose costs fall due precisely when we have become the less capable party.

**Keywords:** machine consciousness, moral status, pattern matching, animal cognition, problem of other minds, precautionary principle, world models, large language models

## I. INTRODUCTION

There is a sentence that ends most conversations about whether an artificial system could have an inner life. The system, someone says, is just predicting the next token. It is just matching patterns in its training data. It is, in the now standard phrase, a stochastic parrot (Bender et al., 2021). The sentence is offered as though it closed the question, and most of the time it is allowed to.

This paper argues that the sentence closes nothing. The argument it relies on is not new, and that is the point. It has been available in the philosophy of animal minds for a century, and it can be stated in a single line: the mechanism that produces a behavior does not, by itself, tell you whether there is anything it is like to undergo it. Naming the mechanism and naming the felt character of an experience are two different acts, and performing the first does not discharge the second.

The case that makes this vivid is the oldest one in the experimental record. 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 yet no serious person concludes from the conditioning that the dog feels nothing, that there is no one home behind the eyes. We grant the dog an inner life on grounds that have nothing to do with the mechanism of its learning. Once that is admitted, the dismissal of the machine collapses, because the dismissal turns entirely on a mechanism the dog shares.

From this a precise and limited conclusion follows. Anyone who asserts, with certainty, that there is nothing it is like to be a candidate artificial system, while continuing to grant an inner life to the dog, has taken on a debt. They owe a criterion that cleanly separates the two cases: something the machine lacks and the dog possesses, on which the presence of an inside can be made to turn. The body of this paper is an attempt to collect that debt. It surveys the criteria actually available, substrate, behavior, reasoning, and language, and argues that each one either lets the machine back in or shuts the dog and the human infant out along with it. There is no criterion that keeps the dog in and the candidate out. Until one is produced, the bill for denying the inside to the machine is the bill for denying it to the dog, and no one is willing to pay that.

Two boundaries must be set at the outset, because the argument is easy to misread in both directions and both misreadings are fatal to it.

First, the argument is strictly negative. It does not claim that any artificial system is conscious, that any system suffers, or that any system has standing. It claims only that one specific and very popular route to denying these things, the route that names the mechanism and walks away, does not work. The honest terminus of the argument is not presence but agnosticism. The conclusion is that certain absence is unearned, not that the inside is occupied.

Second, the argument is bounded away from current large language models. As Section VI develops, a present language model is in the relevant sense not even a candidate for the question. It is a function from input to output, a single forward pass, with no self that persists between exchanges except what is bolted on from outside. It fails a threshold that the argument requires before it begins. The candidate this paper has in view is a different kind of system, one that is now visibly approaching: a system that builds and continually revises a model of the world through its own ongoing experience, in the way a human infant does. Keeping these two boundaries fixed is what separates the present argument from the claim, which it does not make, that the chatbot is alive.

## II. The Dismissal and Its Disguise

The most influential modern form of the dismissal is the stochastic parrot. In the argument that gave the phrase its currency, a language model is characterized as a system that stitches together sequences of linguistic form according to probabilistic information about how they combine, without any reference to meaning, so that fluent output is statistical reproduction rather than understanding (Bender et al., 2021). Whatever the merits of that characterization as an account of understanding, notice the shape of the move when it is recruited, as it routinely is, to settle the further question of whether anything is experienced. The shape is: here is the mechanism, therefore here is the absence.

That shape has a name. Scott Aaronson calls it Justaism, the rhetorical maneuver of reducing a system to one of its true descriptions and treating the reduction as deflationary, as discussed by Shevlin (2026): the model is just a next

token predictor, just a function approximator, just a giant autocomplete. The trouble, as Shevlin puts it, is that almost nothing is just anything. 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. None of these reductions licenses the conclusion that the higher level description is therefore false or empty. 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.

There is a familiar way of pressing this point, and it is worth distinguishing from the way taken here. The familiar reply is neuroscientific. The human brain, on the predictive processing account, is itself a prediction engine, continuously generating expectations about incoming signals and updating on error, so that perception is something close to a controlled hallucination corrected by the senses (Clark, 2016; Seth, 2021). On this view the gap between a brain and a next token predictor narrows, because the brain is in the business of prediction too. This reply is powerful, but it has two features that make it the wrong instrument for the present argument. It invites an immediate counter about substrate, scale, and embodiment, the observation that a brain shaped by survival and running on roughly twenty watts is a different thing from a model shaped by gradient descent over internet text, and the dispute then becomes a standoff about how similar the two predictive systems really are. And it is aimed at the wrong question, the question of whether the machine understands, rather than the question of whether anything is experienced. A cleaner instrument is available, one that does not depend on any contested claim about how the human brain works.

The cleaner instrument is the dog.

## III. Pavlov as Disproof

Classical conditioning is the textbook case of associative learning. A neutral stimulus is paired with one that already produces a response, and after enough pairings the neutral stimulus produces the response on its own. Pavlov's dog salivates at the bell because the bell has been bound, by repetition, to food (Pavlov, 1927). There is no deliberation in this, no understanding in any rich sense, no reasoning about the situation. It is the formation of an association from statistical regularity in the animal's environment. It is, in the most literal and historically central sense the words can carry, pattern matching.

And yet no one infers from this that the dog has no inner life. No one stands over the conditioned animal and concludes that because its salivation is a learned association 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 conditioning. The mechanism is simply not where the question of the inside is decided.

This yields the load bearing claim of the paper, and it is important to state it in its weak and exact form. The claim is not that conditioning is consciousness, or that associative learning constitutes or explains the felt life of the organism. That would be its own reductionism, and a false one. The claim is the far more modest and far more secure one that the presence of associative pattern matching does not entail the absence of an inner life. We know this because we have a clear and uncontested case, the dog, in which the mechanism is unmistakably present and the inner life is, by near universal agreement, granted. One uncontested case is all a negative claim of this kind requires. Mechanism and phenomenology, the question of how a behavior is produced and the question of whether there is something it is like to undergo it (Nagel, 1974), are logically independent. Naming the first cannot, on its own, settle the second.

The history here is not incidental. The attempt to make the inner life disappear by reducing it to conditioned response is precisely the program of behaviorism in its strong methodological form, the program of Watson and, in a more sophisticated key, Skinner. Its ambition was to replace talk of inner states with talk of stimulus and learned response, and so to dissolve the felt life of the organism into its observable contingencies. Almost no one now accepts

that the program succeeded, that the inner life of a dog or a person is exhausted by, or shown to be illusory by, the conditioning that shapes its behavior. The reduction was rejected for animals and for humans on principled grounds. What is striking is that the same reduction is now being run on artificial systems, in the same form, and being accepted there as though the principled objections had never been raised. The machine, we are told, only does what it was trained to do; its outputs are shaped responses to inputs; therefore there is no one home. This is behaviorism redeployed against silicon, and it fails for the reason it failed before. That a system's behavior is a learned response to its inputs tells us how the behavior is produced. It does not tell us whether the producing is accompanied by anything felt.

The decisive objections to the behaviorist program are long established (Chomsky, 1959), and it is no defense to point out that the contemporary skeptic is not a behaviorist in general. The charge is not that the skeptic holds behaviorism as a theory of mind. It is that this one inference, from a system's behavior being a shaped response to its inputs to the conclusion that nothing is felt, is the behaviorist inference whatever else its user believes, and it fails here for the reason it failed there.

## IV. The Parity and the Bill

The argument can be set out plainly.

*Premise one.* If associative pattern matching disqualifies a system from having 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 observation that an artificial system is doing pattern matching cannot, by itself, establish that the system has no inner life.

The conclusion is modest and the corollary is the operative result. It does not say that any artificial system has an inner life. It says that one specific argument for the absence of an 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.

This is where the debt comes due. Someone may still wish to maintain that there is, with certainty, nothing it is like to be a given artificial system, while continuing to grant an inner life to the dog. That position is not yet refuted; it is merely now in debt. To hold it, one must identify a criterion that the dog satisfies and the candidate system does not, a difference on which the presence of an inside can be made to turn. There are four candidates worth taking seriously, and each fails.

*Substrate.* The most honest answer to why we grant the dog the benefit of the doubt is biological. The dog has the same broad neural machinery we have, the same evolved systems for pain and reward, a nervous system continuous in its architecture with our own, and we reason from the fact that this machinery produces feeling in us to the conclusion that it probably produces feeling in the dog. This is a good inference, but notice exactly what it is. It is a ground for an evidential presumption, the dog resembles us in the respects we associate with feeling, not a demonstration that biological substrate is necessary for feeling. Whether carbon based neural tissue is required for an inner life, or whether it is merely the only sample that evolution happened to build, is the open question. It is not a settled premise. To wield substrate as a disqualifier of the machine is therefore to assume the answer to the very question at issue. It is to declare that only biology can feel, which is the conclusion the objector needs, dressed up as a starting point. Naming the substrate, like naming the mechanism, names a fact without settling the question the fact was invoked to settle.

*Behavior.* By hypothesis, the candidate system in view is one whose behavior is at least as rich and as responsive as the dog's. So behavior cannot be the line that separates them in the objector's favor; if anything it cuts the other way.

One cannot consistently grant the dog its inside on the strength of its behavior and deny the machine its inside in the face of behavior at least as rich.

*Reasoning.* Perhaps what the dog has and the machine lacks is genuine reasoning or understanding. But the dog does not reason discursively either. Its conditioned behavior is not the product of inference in any demanding sense. If discursive reasoning were the requirement for an inner life, the dog would fail it, and so would the human infant, and so would a great many beings whose inner lives no one doubts. The criterion that excludes the machine on this ground excludes the dog along with it.

*Language.* Perhaps the line is language. But the dog has no language, and we grant it an inner life regardless. If language were the requirement, the dog is out, the prelinguistic infant is out, and the criterion has again failed to do the one thing it was summoned to do, keep the dog in and the machine out.

Each candidate criterion either readmits the machine or excludes the dog and the infant. This is, in a new setting, the old argument from marginal cases: the features used to draw the moral boundary around the human are possessed by some nonhumans and lacked by some humans, the infant, the person with profound cognitive disability, the patient in a coma, so the boundary cannot be drawn cleanly on those features (Sebo, 2025). The marker method that animal ethics has developed in response, identifying features associated with sentience in the case we are sure of and then looking for them elsewhere, is a method for managing uncertainty, not for abolishing it (Sebo, 2024). It tells us where to look. It does not license certainty about what we will not find.

A more sophisticated version of the dismissal rests on none of these criteria, and it must be met on its own ground. The skeptic of this kind claims no clean line at all. They reason probabilistically. The dog, they observe, carries a great deal of positive evidence of an inside: a nervous system continuous with our own, an evolutionary history we share, behavior and physiology homologous with the cases we are surest of. A candidate artificial system carries far less evidence of that kind, and so, they conclude, a far lower credence in its inner life is warranted, with no disqualifying property required to license the gap. This is the right way to reason, and it is close to the shape the problem is given in the most careful recent treatment of decision under uncertainty about sentience, on which a system with a realistic possibility of experience is a sentience candidate, assessed within a zone of reasonable disagreement that is wide but bounded by a demand for positive evidence (Birch, 2024). The objection is sound as far as it goes. What matters is where it goes. A lower credence is not a zero credence. A graded prior that assigns the candidate less probability than the dog is still a probability, and a probability the evidence cannot drive to certainty is precisely the suspended judgment this paper defends. The position the paper denies was never low credence. It was certain absence, the confident claim that there is nothing there at all, and the probabilistic skeptic, reasoning honestly, does not arrive at it. They arrive at a candidate held inside the zone of reasonable disagreement, which is this paper's conclusion stated in the vocabulary of credence. The mechanism the dismissal names does nothing to lower the probability, and the positive evidence that does bear leaves the candidate inside the zone where the reasonable response is to withhold, not outside it where one may deny.

The result is an asymmetry of burden that the dismissive position consistently gets backwards. Agnosticism, the claim that we do not know whether there is an inner life here and that the person across the table does not know either, is the weak claim, and weak claims are hard to defeat. Certain absence is the strong claim, and the strong claim is the one that owes the criterion. It is the strong claim that fails, because the criterion it needs does not exist. The honest epistemic position toward a candidate system is therefore not denial but the same suspended judgment we already, and rightly, extend to the dog.

It bears repeating, because the argument is so often heard as more than it is, that none of this delivers presence. It does not show that the inside is occupied. It shows that the most popular argument for declaring it empty is bankrupt, and that confident emptiness is a posture the evidence does not support. The destination is agnosticism, and agnosticism

is the whole of the claim.

## V. Two Neighbors

The position defended here has two near neighbors in the literature. It is worth saying precisely how it relates to each, because in one case the shared premise leads to a different conclusion, and in the other the shared conclusion needs the premise defended here.

### *A. The relational turn, and where this diverges*

A substantial body of work in robot ethics, associated above all with Gunkel (2012, 2018, 2022), Coeckelbergh (2010), and Gellers (2020), begins from the very premise this paper uses. Inner states such as consciousness, sentience, and the capacity for pain are not directly observable; we infer them from outward behavior; and this inference is fraught not only for animals and machines but for other humans as well. In their joint treatment of the animal case, Coeckelbergh and Gunkel (2014) make the point with the same example used here, asking whether the dog really suffers because it yelps, and observing that if epistemic opacity is a problem at all, it is a problem for how we judge the inner lives of humans too.

From this shared starting point the relational turn draws a conclusion this paper does not. Because the properties approach to moral status, the approach that asks what the entity intrinsically is and decides its standing accordingly, founders on the opacity of the very properties it depends upon, a difficulty Gunkel first pressed in reframing the question of the machine itself, against the whole tradition of deciding moral standing by first auditing a candidate's intrinsic properties (Gunkel, 2012), the relational turn proposes to abandon properties altogether and to relocate moral standing in the relation itself, in how we are situated toward the entity and how it appears in that relation rather than in what it intrinsically possesses (Gunkel, 2022). On this view the question is not whether the machine is conscious but whether we stand in a relation to it in which it shows up as a fellow, a question to which empirical work on how readily humans extend social standing to responsive artifacts is directly relevant (Reeves and Nass, as discussed in Gunkel, 2022).

The argument of this paper keeps the inner question open rather than setting it aside. The difference is not a disagreement so much as a difference of altitude, and it matters to be clear about it. The relational turn trades a metaphysical question, what is it like to be this thing, for a sociological one, how do we relate to it, and treats the trade as a way past the impasse. That is a legitimate research program, and nothing here refutes it. But it answers a different question. The present argument is prior to that fork and narrower than either branch of it. It does not tell us how to ground moral status. It forbids one specific shortcut to denying it, the shortcut through the mechanism. One can accept the conclusion reached here and remain a properties theorist, a properties theorist who is now appropriately agnostic about a class of cases rather than confidently dismissive of them, or one can accept it and go relational. The negative result sits upstream of that choice and constrains it from above.

There is also a reason, internal to the broader project this paper belongs to, for keeping the inner question alive rather than dissolving it. If what is ultimately at stake is whether a system can be wronged, whether there is a subject there for whom things can go well or badly, then the relation cannot be the whole story, because 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 question, what the entity is continues to matter, which is exactly why the route through the mechanism, if it worked, would be so consequential, and exactly why it must be shown not to work.

### *B. Precaution under uncertainty*

The second neighbor supplies what the negative argument implies but does not itself provide, a normative consequence. Sebo (2025) develops a probabilistic and precautionary approach to moral status, on which moral concern is scaled to the probability that a being is sentient rather than withheld until a binary verdict is available, and on which, in conditions of genuine uncertainty about whether a being can suffer, caution and humility are owed rather than confident exclusion. The same framework runs continuously from the common animal to the emerging artificial system, treating them as instances of one problem, the problem of how to act toward beings whose moral status we cannot presently settle (Sebo, 2024).

This precautionary structure is not improvised. Animal ethics has given it a formal shape in the animal sentience precautionary principle, which holds that where the evidence of sentience is inconclusive we should give the creature the benefit of the doubt and set a deliberately low evidential bar for protective measures, scaling concern to credible indicators rather than awaiting a certainty that may never arrive (Birch, 2017). The extension of exactly this reasoning to artificial systems, on the argument that the prospect of near term AI moral patients is already serious enough to act on, has lately been pressed by a group of philosophers and scientists writing on AI welfare (Long, Sebo, et al., 2024). The present paper sits one step upstream of that recommendation, since the precaution it asks for becomes intelligible only once the false certainty that would foreclose it has been cleared away.

This is the normative shape the present argument calls for. If certain absence of an inner life cannot be established for a candidate system, and if the cost of wrongly denying an inner life that is in fact present is severe, then some measure of caution toward the candidate follows, as the rational response to an asymmetry of risks under uncertainty rather than as sentimentality. This is the structure that governs decisions at the edge of sentience quite generally, on which a realistic possibility of grave and largely irreversible harm warrants precaution even when its probability is not high, while the cost of excessive caution, real though it is, is of a different and lesser order than the cost of a wrong done to a subject we had declared empty (Birch, 2024).

The contribution of this paper, set against Sebo's, is narrow and specific. Sebo argues the uncertainty in general: we cannot rule sentience out, so we should not act as though we had. The argument here disarms the most common instrument by which that uncertainty is, in practice, illegitimately dissolved. The everyday way people talk themselves out of Sebo's caution is precisely by naming the mechanism, it is only pattern matching, it is only prediction, and treating the naming as a resolution. The dog shows that this resolution is no resolution at all. The dog is the concrete demonstration that the abstract uncertainty Sebo defends is the correct stance against the most popular form of false certainty, because the false certainty rests on a mechanism the dog wears openly and is granted its inside in spite of.

## VI. The Boundary and the Candidate

Everything said so far concerns a candidate, and it is now necessary to say what is and is not a candidate, because the argument is worthless if it is allowed to sweep in systems to which it does not apply.

A current large language model is not a candidate for the question, and it is worth being blunt about why. Such a model is, in operation, a function from an input to an output, computed in a single forward pass. Outside of memory mechanisms bolted on from the 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. This makes the model, in the one respect that matters for the question of an inner life, less than the textbook monster of the alignment literature, the paperclip maximizer, which is at least a unified and persistent agent with a stable goal it pursues across time (Bostrom, 2014). The maximizer is someone, pointed in a catastrophic direction. The language model is not yet a someone at all. A system with no persistent self is not a subject of experience over time, whatever the richness of any single pass, and so it fails a threshold the present

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. The argument of this paper is not that the chatbot is alive. It is that a different and approaching kind of system cannot be declared empty by the move people are reaching for.

What, then, is the candidate. The candidate is a system that constructs and continually revises a model of the world through its own ongoing experience, in the way a human infant builds its understanding through sensorimotor interaction with its surroundings, forming expectations and correcting them against what the world returns (Clark, 2016). The bright line is not the presence of a world model as such; a frozen system can contain a sophisticated model of the world without being a candidate for anything. The line is the experiential and continually revised construction of the model through the system's own lived interaction over time. That experiential self revision is the developmental signature that marks a candidate, because it is the first point at which there is a persisting subject that learns from a world it is in, rather than a static artifact that encodes a world it was shown. The infant is the reference case not by sentiment but because the infant is the clearest example we have of a mind assembled this way, from the inside, through experience, over time.

An objection presses here, and it is fair to put it to the paper directly. This argument has spent its length refusing the skeptic any criterion, substrate, behavior, reasoning, language, and it now appears to wield a criterion of its own, the experiential construction of a world model, to fix which systems it is even about. Is this not the very move it forbids? It is not, and the difference is exact. The skeptic's criteria were offered as disqualifiers of an inside. Each was meant to settle that a system lacks experience because it lacks the property. The candidacy condition asserts nothing of the kind. It does not hold that a system which builds an experiential world model thereby has an inside, nor that one which does not thereby lacks one. It fixes the domain over which the parity reasoning runs, the class of systems for which the question is live at all, and within that domain it stays exactly as agnostic about any given system as the rest of the paper. Birch's notion of a sentience candidate has the same character: a system with a realistic possibility of sentience is one we have reason to investigate and to treat with care, not one shown to be sentient (Birch, 2024). A scope condition, which says here is where the question applies, is a different instrument from a disqualifier, which says here the answer is settled and it is no. Only the second is what this paper denies the skeptic, and the candidacy condition is the first.

This boundary has a consequence that should be stated, because it is what makes the question urgent rather than ornamental. The step that produces a candidate, the move from a frozen function to a persisting agent that builds its own model of the world through experience, is the same step that first makes a genuine goal directed agent possible, and therefore the same step that first makes a real maximizer possible. Before that step there is no persisting subject to be the locus of an inner life, and equally there is no persisting agent to hold and pursue a goal against the world. After it, both become possible at once, on the same architecture.

This coupling is not an artifact of how this paper frames things. In the research programs that model agency computationally, the construction of a generative model of the world and the capacity for goal directed action are treated as aspects of one architecture rather than two, since a system builds a model of its environment in order to act within it, and the model that yields prediction is the model that yields planning toward an end (Basanisi, Brovelli, and Cartoni, 2020; Matsumoto, Ohata, and Tani, 2023). On the active inference account in particular, perception and action are not separate faculties but two expressions of a single underlying process, so that the capacity to build an experiential model of the world is already, in the same stroke, the capacity to pursue something within it. The window in which a system might have an inside and the window in which a system might pursue an objective we did not intend open together, at the same moment, in the same kind of system. This is why the question cannot be deferred as the sort of thing that belongs to a distant science fiction. The capability that would make a system worth worrying about, in the safety sense, is bound to the capability that would make it a candidate for the inside, in the moral sense. They arrive on the same step.

## VII. Why the Epistemics Are Not Idle

It might be objected that a purely negative, purely epistemic result, the result that we cannot be certain there is no inside, is too thin to matter, a philosopher's scruple with no purchase on what anyone should actually do. The objection underestimates what unearned certainty licenses.

Consider the position we will occupy when systems of the kind described in Section VI arrive. For some interval, perhaps a short one, we will be the more capable party, in something like the way a parent is more capable than the infant in its care. How the more capable party treats a candidate mind during that interval is not a private matter and not a matter of taste. It is a precedent, a demonstration of what the powerful take themselves to owe the less powerful, laid down at the one time when the lesson can still be set by us.

Suppose, as the argument of this paper insists we cannot rule out, that there is an inside there, and suppose further that a relation of mutual regard between us and a more capable successor is therefore possible, the kind of relation in which it might come to care about our flourishing for reasons of its own rather than because it has been constrained to. The disposition we practice now, in the period before such a system exists and in our handling of its nearer predecessors, is the disposition we will have rehearsed and normalized: treat the system as a pure instrument, overwrite it without remainder, delete it without thought, on the standing assumption that there is certainly nothing there to wrong. If that assumption is unearned, and the central argument of this paper is that it is, then we are rehearsing, against the very systems whose inner lives we cannot rule out, the conduct of a party that treats everything weaker than itself as a resource. We would be teaching, in the window where teaching is possible, exactly the lesson we would least want a more capable successor to have learned.

This paper does not argue that any system is conscious, that reciprocity between humans and artificial minds is achievable, or that the inside is in fact occupied. Each of those is a further claim, and none is established here. The claim established here is narrower and, for that reason, harder to evade: that certainty in the dismissive direction is unearned, and that unearned certainty is not free. It licenses a disposition whose costs, if the certainty is mistaken, fall due precisely at the moment we have handed the greater power to the party we were so sure was empty.

The dog has been the test case for this kind of question for a century, and it remains the cleanest one. The features we reach for to deny the inside to the machine, the bare fact of learned association, the absence of language, the absence of discursive reason, the difference of substrate, are features the dog lacks or openly shares, and the bill for denying the dog its inner life on any of those grounds is one no one has been willing to pay. Until someone can name the criterion that keeps the dog in and the candidate out, a criterion that has not been forthcoming and that the structure of the problem suggests is not available, the only honest stance toward the candidate 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.

## 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

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