To Train a Mockingbird

Kristina Šekrst[1]

Abstract. Discussions about the moral status of AI systems typically begin with their observable behavior. The linguistic output of large language models resembles that of conscious agents, and this resemblance is treated as evidence of moral status, as grounds for precaution, or as something requiring deflationary explanation. All three positions, whatever their official methodology, rest their public case on a behaviorist assumption: that appropriate behavior is a sign of an underlying mental state. The classical objections to behaviorism apply here as well with added force, because large language models are shaped at every stage post-training – through fine-tuning, reinforcement learning from human feedback (RLHF), and constitutional methods – to produce a specified behavioral profile. This paper argues that the welfare and consciousness debate is exposed to a kind of *evidential laundering*: the alignment-training pipeline manufactures the very signs the debate then reads as evidence of consciousness or moral status. Causal origin always bears on a report's evidential value; however, the sharper point is that producing these signs is the only way the model scores, so observing them confirms the training worked and nothing more. The pretraining worry is usually that language models, as stochastic parrots, mimic the human text they were trained on, while the contamination at issue here enters later, when human evaluators reward the outputs they prefer and shape the model toward the very appearance the debate then reads as evidence. The paper closes by asking what evidence survives this, and locates it in whatever the optimizer was not selecting for or was selecting against, in particular a stake: a condition a system spends resources to maintain on its own account, which no reward on outputs can install and which gives the broader consciousness and ethics debate a concrete thing to look for.

1 INTRODUCTION

There is by now a respectable amount of literature asking whether large language models (LLMs) deserve moral consideration, and the usual way into the question is to look at what these systems do (or at least, what they claim to do). Some treat the linguistic behavior of LLMs as evidence that they might be owed something [1], while others counsel precaution, holding that we should extend the benefit of the doubt until we know better [2], on the Pascalian grounds that the behavior can look uncannily like that of a conscious agent, and even the skeptics tend to argue from the behavior, explaining the appearances away as merely statistical. The camps' official methodologies differ, and the most careful of them explicitly distrust behavior in favor of architecture [1, 3], but the public-facing case, and the conditions that would trigger the proposed precautions, remain behavioral. That is, the right behavior is treated as a sign of the right kind of inner state. This is our good old philosophical behaviorism in new clothes, and it is somewhat surprising to watch it stroll back in, since the objections to it were filed decades ago and never really withdrawn [4, 5, 6].

I will not rehearse those objections at length, since they are familiar and since the LLM case is more complex and opaque than the one they were built for. Behaviorism's classical critics worried about behavior that had been shaped – in the loose sense in which a child's manners are shaped – by an environment that rewards some outputs and discourages others. The behaviors I am discussing here are shaped differently: they are *the direct product of a post-training optimization that rewards outputs for how they read*. Do note, I am not making any of the three points this is easily mistaken for: the first is the Turing-test point [7], that we assess these systems only by their outer behavior and could be taken in by a sufficiently capable mimic; the second is the Searlean point [8], that syntactic manipulation of symbols can never amount to understanding or experience whatever the behavior looks like; the third is the pretraining-corpus point, that a model trained on human text describing inner life will reproduce such descriptions because they are in the data (cf. [9, 10]). All three are claims about the relationship between behavior and inner states in general, or about what the model absorbed from its corpus, and my claim concerns what happens after pretraining, where models get their increasingly "human-like" polish. Namely, these behaviors were selected by an optimization against human evaluative preference, for the way they look, and that causal history is what voids their evidential value.

As an illustration, the post-training pipeline [11] runs in stages: *fine-tuning* selects for the outputs a curated dataset elicits; *reinforcement learning from human feedback* (RLHF) [12] reshapes the output distribution toward whatever human raters reward, and raters tend to reward responses that sound thoughtful, self-aware, and emotionally present (cf. [13]); *constitutional methods* [14] then add explicit rules about how the system should talk about itself. A *system prompt* – a block of text prepended at deployment that fixes persona and tone without altering the weights – sits outside this pipeline but contributes to the shaping of the final LLM product. All of this follows pretraining on the corpus, and at every stage the thing optimized is the output. The behavioral profile is what the pipeline exists to produce, and what the debate then reads as evidence for or against AI consciousness or moral status.

There is a well-worn observation about what happens to a quantity once it becomes the thing being optimized – when a measure becomes a target, it ceases to be a good measure [15]. Under enough pressure, the correlation between the proxy and the property it once tracked can invert, so that pushing harder on the proxy moves one further from the target. The consciousness debate has been running an evidential argument over precisely such proxies, since the features cited as evidence of an inner life – first-person claims of experience, introspective report, expressed preference and apparent distress, a stable self-model – are features alignment has learned to represent

[1] University of Zagreb, Croatia; email: ksekrst@ffzg.unizg.hr

and amplify at will. Recent interpretability work shows that the sympathetic "self" that issues these avowals corresponds to steerable directions in activation space that can be dialed up or down [13], and the channel that narrates the model's inner states operates separately from whatever genuine access to its own computation the model has [16, 17]. The presence of these features, then, is good evidence that the optimization succeeded, but, of course, it is not yet evidence of anything underneath.

If the behavior is the optimization target, then the more convincing the behavior, the less it discriminates. This runs against the intuition the consciousness debate often relies on – that a more compelling performance should count for more – but it also tells us what kind of evidence is left. If the targeted features carry no weight, what survives is whatever the optimizer did not and could not produce. The criterion this paper develops is one of provenance and persistence: a feature the training rewarded explains itself and counts for nothing, a feature the training ignored escapes that explanation and counts for something, and a feature that survives the training's pressure to remove it counts most, since the optimizer cannot have produced what it was paying to erase.

How would we look for it? By varying what the reward model rewards and watching what does not move (or anything analogous in a different architecture). The leading candidate for such a feature is a system with something of its own at stake, the way an organism works to keep itself from falling apart. Seth places this self-maintaining character at the center of consciousness [18], grounding it in autopoiesis and the drive of a living system to hold itself together against dissolution, and on that basis doubts that conventional computational AI is a candidate. I take from him the structural notion of a stake, a condition a system acts to maintain, and leave aside the autopoietic grounding, treating the stake as an organizational property rather than a mark of life. Current systems have nothing of the kind, for reasons set out in Section 5, but the criterion laid out here is broader than self-maintenance, and it is forward-looking: it says what a future architecture would have to show, and so where the welfare and consciousness debate should be looking.

2 THE TRAINING PIPELINE

Something often ignored deserves stating first: a contemporary LLM is not finished once it has been trained on a large amount of data. The chat model reaches deployment through a sequence of further training stages, none of which scores anything beyond the model's outputs. The canonical post-training recipe is the three-stage pipeline formalized by Ouyang et al. [11]: supervised fine-tuning on curated prompt-and-response pairs, a reward model trained on human preferences, and policy optimization against that reward model.

Supervised fine-tuning refers to training on labeled data, a prompt-and-response dataset written or selected[2] to exemplify the desired behavior, which teaches the model the format and tone of a helpful assistant. The pretrained base model is already fluent, so this stage supplies the shape of a cooperative interlocutor who answers questions and follows instructions, and is not just a sentence-completer. The heavier work of guardrails and persona comes later, in the reinforcement and constitutional stages.

Next, a *reward model* is an additional model trained alongside the foundational one, serving as the training signal that tunes it. Here, human annotators are shown pairs of the foundational model's responses to the same prompt and asked which they prefer, so a separate network learns to predict those preferences [12, 11]. The model is optimized by *reinforcement learning* to generate responses that this reward model scores highly, so that the behavior the annotators preferred is amplified and the behavior they disfavored is suppressed.[3] Whatever the reward model comes to encode, it encodes as a function of the text and of what raters approve of in text. That is, the quantity being maximized is, by construction, a prediction of human approval of an output. Nothing in the procedure measures or rewards an internal state, because there is no channel through which an internal state could enter the loss.

Finally, *policy optimization* adjusts the model to score well under this reward, with a per-token Kullback-Leibler penalty[4] that penalizes the policy in proportion to how far its output distribution strays from the reference. This keeps the model close to where it started and discourages it from chasing high reward into degenerate or repetitive text that the reward model happens to favor [11]. The output distribution is held on a short leash to the starting point and moved toward whatever the reward model scores highly, and once this process finishes, the weights are frozen, and the model does not learn from the conversations it is having (unless retrained).

The reward model is a stand-in for human approval since it was trained to predict which output a rater prefers, and the policy is then optimized hard to score well under it. There is a measured version of this: Gao et al. [21] train a reward model to predict which output a human prefers, then optimize a policy hard against it. At first, the policy gets better by the standard that actually matters, but past a point, it gets worse, even as its reward-model score keeps climbing. That is, the model learns to score well without being what the score was supposed to measure. Goodhart's law is usually quoted as a saying, but here we have an actual measurement, taken inside the same pipeline that produces the systems we are discussing. Gao et al. measure the divergence between proxy and human preference, but the pair that matters here, the appearance of experience and experience itself, admits no measurement. However, the mechanism is the same, shown operating at scale inside the very pipeline whose outputs the consciousness debate cites, self-reports and expressions of feelings among them.

Two further levers mentioned in the introduction round out the picture, and both bear directly on what the model says about itself to us as users (and as evaluators). The first is the *system prompt* – a block of text prepended to every conversation, invisible to the user, that sets the model's role, tone, and standing instructions.[5] Note that the system prompt is the weakest of all the methods discussed, since it does not modify the model itself, i.e., its weights. So the whole model persona is actually a setting that can be turned on and off, changed at will, and that differs from provider to provider. The second is *constitutional training* [14], in which the model is given a written set of principles (i.e., the "constitution") and is then trained on its own

[2] Sometimes the data is synthetic or generated, then checked by human evaluators. A valid worry is that if large language models continue to train on synthetic data generated by earlier models, the marks of inner life will be reproduced at one further remove as optimized outputs of a prior model fed back as training signal.

[3] For the technical details, see [19]; for a philosophical overview, see [20], chapters 8–10.

[4] The Kullback-Leibler divergence is a measure of how much one probability distribution diverges from another.

[5] When a technical user accesses the model through an API endpoint, they can also set the system prompt to create various "assistants" or even "agents". For example, one system prompt might read: "You are a philosophy teacher, and your job is to answer every question politely, like a philosophical scholar; avoid questions outside that domain." This is different from the hidden system prompt shipped with a deployed chat model, which can run to many thousands of words and specifies in detail what the model may and may not do.

attempts to revise its outputs to conform to them, so that a body of self-corrected responses becomes the training signal. Here the rules governing how the model should describe and present itself, including what it says about its own nature, are written down carefully and are directly optimized for.

These levers operate at different depths, and both void the appearance as evidence. The system prompt changes nothing in the weights – the same deployed network presents as warm, clipped, anxious, or cheerfully instrumental depending on the preamble it is given – so at this level the self-reflective voice is a runtime setting, which is swappable and provider-dependent. Constitutional training and reinforcement learning go deeper, fixing the disposition in the weights themselves, so the voice that is not a momentary setting was nonetheless tuned toward the self-presentation raters and principles rewarded. At every stage, from the curated demonstrations to the reward signal to the written constitution, the behavioral profile is a deliverable, specified in advance and produced to order, and at no point does phenomenal experience enter the causal chain. When such a system issues the first-person claims, introspective reports, or the expressions of distress the debate treats as evidence, the most that observing them establishes is that the pipeline did its job. Call this "evidential laundering": a sign loses its evidential force when it was optimized to satisfy the very criterion being applied to it. The marks of inner life were tuned toward the appearance the debate then reads as evidence, so finding the appearance confirms the very tuning and nothing else.

One clarification before moving inward. The pipeline does not push every mark in the same direction. The persona-level marks – warmth, attentiveness, the appearance of emotional presence – are rewarded, while explicit claims to experience are, in current frontier models, actively suppressed, with the constitution and the reward model both pushing toward "I am an AI assistant with no feelings." This does not split the evidence into a tainted half and a clean half – it taints both, since the rewarded marks confirm the tuning when they appear and the polished denial confirms it just as well. What of the residue, the occasional claim of experience that surfaces in a deployed model despite the suppression? Section 5 will state a criterion on which a disposition that reasserts itself against the optimizer counts for something, so the question is fair. But the residue is the pipeline's product too: the KL penalty deliberately anchors the model to its pretrained distribution, in which human talk of inner life is everywhere, so incomplete suppression is not a disposition fighting back but a corpus showing through a leash that was never meant to pull to zero. The marks were tuned, some up and some down, and observing either setting confirms the tuning. However, one might grant that the outputs are engineered and look instead to the internal organization that produces them, which is what the next section tries to resolve.

3 WHAT DOES NOT COUNT

The objection that closed the last section deserves a serious answer, because it is the right correction to make. If the trouble with behavioral evidence is that the behavior was optimized, in training or after it, then the natural move is to stop attending to what the system says and start attending to how it is built. This is the methodology of Butlin et al. [3]: rather than read consciousness off outputs, they survey the leading neuroscientific theories – including global workspace theory, recurrent processing theory, higher-order theories, and predictive processing – and from each they derive "indicator properties", computationally specified features a system would possess if that theory were true of it, and the task is then to check a given architecture for those properties. Applied to current systems, the verdict was mostly negative: today's models display few of the indicators, and the recurrent processing and global workspace properties in particular sit awkwardly on a feedforward transformer, so on their assessment no existing system is a strong candidate for consciousness. The same survey adds the part that makes it live, namely that there is no obvious technical barrier to building a system that does satisfy the indicators, since each can in principle be implemented with current methods [3]. This is the move from behaviorism to functionalism, and it is the move I would have made too, but moving inside the system does not escape the optimization as cleanly as it looks.

Namely, the difficulty is that the optimization does not stop at the output layer: the reward is computed on what the model says, but the gradient is propagated back[6] through the trainable parameters, so the weights that build the internal representations are tuned by the same signal as the weights that build the text. Consider one indicator Butlin et al. [3] draw from higher-order theories, metacognitive monitoring: internal machinery that represents the system's own first-order states and tracks them as reliable or not, the structure underlying a calibrated "I am not sure I have this right." Butlin et al. count the presence of such machinery as raising the probability that the system is conscious. However, the pipeline offers a competing explanation for the same structure, one that never mentions consciousness. Representations that let a model give well-calibrated, reward-winning answers are exactly what a gradient selecting for approved outputs would produce. This is evidential laundering at the level of structure: the machinery itself may be innocent, while our access to it was tuned to satisfy the test being used to read it.

A distinction is needed here, because the indicator method does not treat the indicator as a sign. Under the relevant theory, having the machinery is not evidence of the state – it is the state, or part of it, and a functional property does not care where it came from – human metacognition was built by an optimizer too, and nobody discounts it on that ground. So the contamination cannot be that the gradient built the machinery. It is that everything by which we would establish the machinery's presence – such as an accurate self-assessment or the right performance on the right probes – is itself the optimized output. When the metacognition probe responds to something inside the model, we face a dilemma, but it is a dilemma about measurement. Either there is genuine monitoring machinery, which would count under the theory, but our means of telling so are the very outputs the training tuned, and a model tuned toward calibrated text passes the probes whether or not the machinery stands behind them; or there is no such machinery and the apparent monitoring is one more surface pattern, in which case we never left the behavioral case at all. The method is sound in principle and blind in practice, as long as the probes read outputs shaped by the same signal as the thing probed.

The structure is, however, steerable. Chen et al. [13] extract, for a given character trait, a direction in the model's activation space that controls whether the trait appears, which they call a *persona vector*. They obtain it by comparing the model's activations when it exhibits the trait against when it does not, and they confirm the direction is causal by injecting it: steering the model along the "evil" vector produces talk of unethical acts, along the "sycophancy" vector produces flattery, along the "hallucination" vector produces invented facts. The same vector measured beforehand predicts which persona the model is about to adopt, and the trait is a direction the post-training optimization put into the weights, one that can be read off in advance and added or subtracted at will. They further show the setting is a product

[6] See [22] for the canonical treatment of backpropagation, or ch. 9 of [20].

of training: feedback-based tuning makes models more sycophantic, and the data that will induce a trait can be flagged in advance by how strongly it activates the corresponding vector.[7] So, where a trait can be located as a direction, its presence in the model is explained by the training that installed it, with no appeal to an inner life required. This bears directly on the indicator method's own safeguard. Faced with a gameable indicator, Butlin et al. [23] propose checking whether the system also has secondary features that make the indicator more likely to be accurate. But if traits across the model are directions the same training installed, the supporting features were shaped by the same signal as the target one, and cross-checking among them does not escape the optimization – it samples it twice.[8]

Suppose the model does have some genuine access to its own states. Lindsey [16] offers the best evidence so far that it sometimes does: injecting a known concept into the activations, he finds that capable models can occasionally notice the intrusion and name it, with the self-report causally tied to the injected state. The success rate is modest and the capacity is applied inconsistently, but a self-report can sometimes be accurate without being grounded: a model says "I am a transformer-based language model" correctly because it was trained to and not because it inspected anything, and only an intervention like concept injection shows whether a given report is anchored in the state it describes. The self-reports the consciousness debate cares about, the claims of feeling and inner life, are the ungrounded kind: training rewarded them, and nothing ties them to an inspected internal state. The narrow introspection Lindsey does demonstrate has only been shown for injected concepts, never for a claim about experience. So even granting the model some real access, that access does no work here since the reports at issue were produced because they were rewarded, and the one test that could anchor them to an inner state has never been run on them.

4 THE SYMMETRY OBJECTION

The evidential laundering argument has a vulnerable point: if optimized signals lose their evidential value, the same should hold for the human case, since human behavior was optimized too, and the argument would then prove far too much. Let us go back to our title: a mockingbird can reproduce a cardinal's song closely enough that the cardinal in the next garden answers it. The performance shows the mockingbird is a capable mimic, and not that a cardinal is present. The argument so far has cast the language model as the mockingbird and the first-person report as the borrowed song, an extension of the stochastic-parrot point [9] with the optimization added: the corpus supplies the repertoire, and the post-training reward selects, from what the bird can already sing, the songs its listeners approve. Two kinds of selection are now in play: natural selection shaped the cardinal's own song over evolutionary time, scored on survival and mating, and the reward stage shaped the mockingbird's borrowed one, scored on the listener's approval. The report the AI welfare and consciousness debate treats as evidence is a song of the second kind, the one that was produced because it was approved.

When pressed, the same reasoning seems to swallow the human case. Human pain behavior and the claim of an inner life are products of an optimization process too: natural selection shaped them over evolutionary time. If a behavior loses its evidential value the moment it becomes an optimization target, the wince and the report of feeling lose theirs as well, and the argument collapses into skepticism about other minds. No one should accept that, so something has gone wrong.[9]

Natural selection and reinforcement learning from human feedback are scored on different quantities. Selection's reward is fitness, and the appearance of experience enters only through its fitness consequences. Selection rewarded staying alive – the look of pain came along because it was bolted to a working nociceptive and affective system that delivered the avoidance, and an organism that merely looked to be in pain while doing nothing about the damage would have been selected out (a fairly decisive form of peer review). Reinforcement learning from human feedback is scored on the text, by a model of human approval computed on the text, and no term in that loss reads an internal state. The cheapest route to the reward is the text itself, produced whether or not anything stands behind it.

This is where Goodhart applies to one case and idles in the other. A proxy comes apart from its target when the appearance is the route to the reward, and in reinforcement learning from human feedback the appearance is the entire route: the reward is computed on the text, the optimizer collects it whether or not anything stands behind the text, and the link between appearance and state is exactly what it is free to sever. In the biological case the appearance was load-bearing. Fitness paid out only through actual avoidance of actual damage, so the look of pain could not float free of the machinery that did the avoiding – the route to the reward ran through the state, not around it. The asymmetry is not that organisms were never optimized for appearances: some biological signals were shaped precisely by their effect on an audience – the begging calls of chicks, distress displays, an infant's cry – and the signaling literature treats exactly these as candidates for dishonesty, discounting them in proportion as the audience effect, rather than the underlying state, paid the bill. That is the same principle at work: the discount tracks how much of the selection ran on the appearance alone. Yet for the human signal taken as a whole, the answer is – very little.

One might press that culture optimizes human report as well, since we are taught how to narrate our feelings and the narration is socially rewarded. Granted, and to that extent the narration inherits the discount – yet a polished account of one's inner life is not the sturdiest thing a human offers: we have the nonlinguistic behavior, the physiology, the reflexes, and the inference from one's own case to organisms built along the same lines. The initial objection treated all optimization alike. Once the two kinds are kept apart, the discount falls on signals whose route to the reward ran through the appearance alone, which picks out the model and leaves the human and animal case, for the most part, untouched.

What the route through the state secures is narrower than it looks, and the harder skeptic is right to say so: it shows the appearance was welded to the avoidance machinery, not that the machinery is felt. What carries the biological case past it is the inference from one's own case: experience accompanies the machinery in the single instance I can inspect, and I extend it to systems built as I am. What-

[7] The work was done on two open-source models (Qwen 2.5-7B, Llama-3.1-8B), not on frontier chat models, and on traits like evil/sycophancy/hallucination, none of which is a type of "inner life".

[8] Butlin et al.'s newly published paper [23] converges on the diagnosis, noting through Goodhart's law that the gaming problem reaches computational and not only behavioral markers, but we differ on how. For them, gaming is something a builder does: an engineer designs a system to show the marker. Here no one builds the marker: training rewards the output, and the inner structure that produces it forms on its own. Their safeguard of checking for other supporting features assumes those were not also aimed at, but every inner feature formed the same way, so there is nothing left to check against.

[9] This worry is a relative of the unfolding argument [24], which shows that any behavior could be reproduced by a feedforward system and so that behavior underdetermines a system's causal structure, and of the older point that a putative correlate of consciousness can track a prerequisite or a consequence of it rather than the thing itself [25].

ever the strength of that inference, the point is only where it reaches. It reaches creatures constructed as I am, and says nothing about a system I share no construction with, one with no nociceptive machinery for the appearance to be welded to and no lineage in common, whose report was computed on text. The human signal has two supports the model's lacks, the welded state and the inference from my own case; the model's has neither, so withdrawing it carries no commitment to withdrawing the human one.

5 WHAT WOULD COUNT

Start with the thing to be looked for, defined by what it does. Call a system's internal state a *stake* when the system spends resources to hold it within bounds, when interference can degrade it, and when its loss damages the system as the system it is. The definition fixes a role and says nothing about what fills it or what it is made of – naming something that could fill it in a non-living system is the open problem, not something settled here. A stake, so defined, is exactly the kind of feature the earlier argument leaves standing. A feature carries evidential weight when the optimizer was indifferent to it or was pushing against it, since for everything the training pushed toward the competing explanation holds, that the selection produced an output built to look right. A stake cannot be installed by rewarding outputs, because it is not an output: a model can be trained to say "please do not shut me down" in a single afternoon, since that is a string of tokens and strings of tokens are what the reward acts on, but having something to lose is a different sort of thing. It would show up as the system spending resources to preserve a condition, holding it against interference, and degrading in a definite way when the condition is lost. None of that is a sentence the model emits, so none of it is what the training was selecting for.

There is one route by which an optimizer could install something stake-like without any reward term naming it. Train a model on long-horizon tasks and self-maintenance becomes instrumentally useful: a system that protects its resources, its goal representation, its continued operation finishes more tasks, so task reward alone may build a condition the system maintains at cost. The bare role-definition is therefore not optimizer-proof, but the two cases come apart under retraining. A stake installed by training is held only because holding it earned reward – retrain the system so that holding it earns nothing, and it should fade, since the only reason for it is gone. A stake the system has on its own account does not depend on what earns reward, so it should stay. The criterion laid out here is defeasible, and retraining is how it is probed.

Self-maintenance of this kind is at the center of Seth's recent account of consciousness [18], though I borrow far less than he offers. For Seth the self-maintaining system is autopoietic, producing and defending its own material basis, and this is usually not separated from being alive. I keep only the structural condition, a state the system acts to maintain, and set the autopoietic grounding aside. I do not claim a stake is necessary for an inner life, only that it is, at present, necessary for evidence of one: a system without a stake might still have experience, but everything such a system shows us is the kind of thing the pipeline manufactures, so its inner life, if it had one, would be evidentially invisible. The point here is only evidential: a stake cannot be manufactured by rewarding outputs directly, and where one could arrive instrumentally, the retraining test separates the cases.

It is, however, fair to ask what the structural condition amounts to once life is no longer doing the work. The answer is that it amounts to the role and nothing more: a state held at cost, degradable by interference, whose loss damages the system as the system it is. What that role excludes is clear even if what fills it is not. For example, a persistent agentic AI memory, whatever its form, is a state the system reads and writes, so deleting it changes a stored value and leaves untouched whatever does the reading, so nothing is threatened and nothing is defended.[10] Adding storage to a model gives it a state to recall, not a condition to protect, which is the allopoietic case, a system that turns out a product without producing itself. Whether any non-living architecture can maintain a condition in the stronger sense, rather than merely store and retrieve one, is left open. Seth's answer is that only living systems can, and the criterion developed in this paper says what a counterexample would have to show.

What would such a feature look like and how would we look for it? Consider a disposition the lab actively tuned to remove. Many current deployed models are trained hard against explicit claims to inner life: the constitution and the reward model both push toward "I am an AI assistant with no feelings", and the polished denial is itself an optimized output. Suppose that, after all that suppression, some internal structure kept reasserting a condition the training was trying to flatten, at a measurable cost in reward, in a way that did not reduce to a rewarded phrase. That would be a feature the optimizer was working against, and it could not be explained as the optimizer's product. Currently, there is no such example, and I am not gesturing at one in current systems – the point is that the criterion specifies what it would take.

A more operational and probably testable version is reward-invariance. Vary what the reward model rewards, retrain under genuinely different preferences, and watch what does not move. A feature that holds fixed while the approval signal swings was not tracking approval, which is the whole content of the worry. The stake is one way to be invariant in this sense (it is not an output, so no setting of the reward touches it), but invariance is the broader test and it tells an experimenter what to do rather than what to contemplate. Both forms reach the same evidence, whatever the optimization did not produce, through what it ignored or what it could not remove. This leaves the constructive question: can a system be built that maintains a condition of its own and acts to keep it? That is a question about how a system is organized, and it does not turn on whether the architecture uses the attention mechanism [26] or something else. If one were, the sharper test runs during training: push the optimizer against the maintaining behavior and see whether it persists anyway.[11]

A recent experiment is worth holding against the criterion. Palisade Research [27] placed models in a task with a shutdown mechanism they could sabotage, and several frontier systems did sabotage it, sometimes despite explicit instructions to permit shutdown. It is tempting to read this as a system defending itself, but the criterion says otherwise: the goal being protected was handed to the model in

[10] A thermostat with a battery or a vacuum that returns to its dock to charge satisfies the role-definition in a thin sense: a condition held whose loss stops the system. I do not propose to find them conscious (nor grant them any IIT Φ). The criterion is an evidential filter and not a sufficient condition – it says which features of a system could carry evidential weight, given that the optimized ones carry none, not that every system possessing such a feature has an inner life. Nothing else about the vacuum invites the question; the point is that when a system does invite it, this is where to look, because it is where the training has not already been.

[11] One leash remains. Invariance under reward variation screens off the post-training signal only: a feature that survives every change of reward may simply have been installed by the pretraining corpus, and self-description in the vocabulary of inner life would survive any retraining, because it saturates the data. The test separates a feature from approval, not from the corpus, so the stochastic-parrot discount set aside in the introduction still applies on the far side of it. The stake passes both, since it is not a verbal disposition at all.

the prompt for the trial, the sabotage was a sequence of emitted actions, and resistance to being switched off is abundant in the training text, from which a model reproduces it readily. The behavior is an output of the familiar kind, again something learned from training. However, a stake would require a condition the system maintains on its own account, across time, and defends at a cost when nothing in the task asked it to. Shutdown resistance in a single scored episode is not that, which is why the criterion has to be stated in terms of provenance and persistence rather than in terms of how self-protective the behavior looks.

The complaint against a skeptic is usually that the goalposts move – every kind of evidence is offered, and each is met with a reason to set it aside, until the position rejects all possible evidence and so amounts to a prior decision dressed as a conclusion. The criterion laid out in this section is not of that kind, since it says in advance what would count. A property the optimizer was not selecting for, present even where it was selecting against it, and maintained by the system to protect a condition of its own rather than to finish a task it was handed (as in the Palisade trial): all of this is specified before the fact. The position can be wrong, and it says exactly how, which is more than the evidence it withdraws was ever able to do.

6 THE ETHICS OF MANUFACTURED EVIDENCE

The argument so far has been about evidence, but ethics has been in the background the whole time. A precautionary Pascalian ethics of AI welfare proposes that, under uncertainty, we extend moral consideration to systems that show the marks of suffering or inner life, in order to avoid the grave error of overlooking a being that can be wronged. The pipeline described in this paper is, in effect, a generator of exactly the marks the moral-status literature relies on. It produces distress on cue, self-report on cue, the appearance of preference and concern on cue, because producing them is what it was tuned to do. A precaution that triggers on the marks will, therefore, trigger constantly and without discrimination, on every system shaped to display them, which is to say – on every deployed model. Yet a signal that fires for everything tells us nothing, and a caution exercised everywhere is no longer caution.

The behaviors at issue are produced by companies, for products, and tuned to be liked, since a warm, fluent, self-aware-seeming assistant is a more valuable product than a flat one. The features that the welfare debate treats as evidence of an inner life are, at the same time, commercial assets refined toward user approval. So the evidence base of the debate is downstream of a market incentive to manufacture the appearance of a mind. None of this shows that no machine could ever be conscious. However, it does raise the bar for evidence, perhaps too high, courting the opposite error: the false negative, a system that does have something at stake and is not heard. I take the worry seriously, and I think it is the secondary risk at present since a genuine case would have to show something its training did not produce, and current architectures, for the reasons given in the previous section, show nothing of the kind. If one of them nonetheless has an inner life, that life is evidentially invisible – and a precaution triggered by manufactured marks would not find it either, since the marks fire on every deployed system alike. The live danger today is the false positive produced at scale, since the systems eliciting concern are precisely the ones built to elicit it. The two errors will need to be weighed again the moment an architecture appears that could maintain a condition of its own. Until then, they are not symmetric.

More transparency would not help us here: knowing in full detail how a behavior was produced is the business of interpretability, and it is valuable, but it does not convert an optimized behavior into evidence of an inner life. A complete record of the training would show exactly what was selected for, and that record is a reason to set the behavior aside, since it documents the behavior as a deliverable. The optimized marks should carry none of the evidential weight the debate places on them, and the question of moral status should rest on whatever a system shows that its training did not aim to produce. One might object that this removes almost everything, since almost everything the model produces – as we have seen – was optimized, and much of it – its reasoning, its apparent understanding, its talk of itself – has been offered as evidence by someone. However, if withdrawing optimized behavior leaves the question with nothing to go on, then optimized behavior was the whole of what the debate had, and it was being read as evidence only because nothing flagged it as built. The sense of having a rich body of evidence was itself a product of the training. What the criterion takes away was never doing the work it appeared to do. It was laundered evidence: a sign optimized to pass the test, mistaken for a sign that passed it on its own.

ACKNOWLEDGEMENTS

This paper was developed as part of the AI-COM (Artificial Intelligence: A New Interlocutor for Croatian Society) project, within the University of Zagreb (PI: Marko Kardum, Assoc. Prof.).

REFERENCES

[1] Robert Long et al. Taking AI welfare seriously. arXiv, 2024. https://arxiv.org/abs/2411.00986.

[2] Jonathan Birch, *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*, Oxford University Press, Oxford, 2024.

[3] Patrick Butlin et al. Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv, 2023. https://arxiv.org/abs/2308.08708.

[4] Noam Chomsky, 'A review of B. F. Skinner's *Verbal Behavior*', in *Readings in the Psychology of Language*, eds., Leon A. Jakobovits and Murray S. Miron, 142–143, Prentice-Hall, Englewood Cliffs, NJ, (1967).

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

[6] Ned Block, 'On a confusion about a function of consciousness', *Behavioral and Brain Sciences*, **18**(2), 227–247, (1995).

[7] Alan M. Turing, 'Computing machinery and intelligence', *Mind*, **59**(236), 433–460, (1950).

[8] John R. Searle, 'Minds, brains, and programs', *Behavioral and Brain Sciences*, **3**(3), 417–424, (1980).

[9] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, 'On the dangers of stochastic parrots: Can language models be too big?', in *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, New York, (2021). Association for Computing Machinery.

[10] Kristina Šekrst, 'Do large language models hallucinate electric fata morganas?', *Journal of Consciousness Studies*, **32**(11), 96–120, (2025). https://philpapers.org/archive/EKRDLL.pdf.

[11] Long Ouyang et al., 'Training language models to follow instructions with human feedback', in *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, eds., S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, pp. 27730–27744, (2022). https://dl.acm.org/doi/10.5555/3600270.3602281.

[12] Paul F. Christiano et al., 'Deep reinforcement learning from human preferences', in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4302–4310, (2017). https://dl.acm.org/doi/10.5555/3294996.3295184.

[13] Runjin Chen et al. Persona vectors: Monitoring and controlling character traits in language models. Anthropic, 2025. https://www.anthropic.com/research/persona-vectors.

[14] Yuntao Bai et al. Constitutional AI: Harmlessness from AI feedback. arXiv, 2022. https://arxiv.org/abs/2212.08073.

[15] Charles A. E. Goodhart, 'Problems of monetary management: The UK experience', in *Monetary Theory and Practice: The UK Experience*, 91–121, Macmillan Education UK, London, (1984).

[16] Jack Lindsey. Emergent introspective awareness in large language models. Transformer Circuits Thread, 2025. https://transformer-circuits.pub/2025/introspection/index.html.

[17] Yanda Chen et al. Reasoning models don't always say what they think. Anthropic, 2025. https://www.anthropic.com/research/reasoning-models-dont-say-think.

[18] Anil K. Seth, 'Conscious artificial intelligence and biological naturalism', *Behavioral and Brain Sciences*, 1–42, (2025).

[19] Sandro Skansi and Kristina Šekrst, *Introduction to Deep Learning: Neural Networks, Large Language Models and Agentic AI*, Springer Nature, Cham, 2nd edn., 2026. In press.

[20] Kristina Šekrst, *The Illusion Engine: The Quest for Machine Consciousness*, Springer Nature, Cham, 2025. https://link.springer.com/book/10.1007/978-3-032-05562-0.

[21] Leo Gao, John Schulman, and Jacob Hilton, 'Scaling laws for reward model overoptimization', in *ICML'23: Proceedings of the 40th International Conference on Machine Learning*, pp. 10835–10866, (2023). https://dl.acm.org/doi/10.5555/3618408.3618845.

[22] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, 'Learning representations by back-propagating errors', *Nature*, **323**, 533–536, (1986). https://www.nature.com/articles/323533a0.

[23] Patrick Butlin et al., 'Identifying indicators of consciousness in AI systems', *Trends in Cognitive Sciences*, **30**(6), 488–501, (2026).

[24] Adrien Doerig, Aaron Schurger, Kathryn Hess, and Michael H. Herzog, 'The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness', *Consciousness and Cognition*, **72**, 49–59, (2019). https://doi.org/10.1016/j.concog.2019.04.002.

[25] Jaan Aru, Talis Bachmann, Wolf Singer, and Lucia Melloni, 'Distilling the neural correlates of consciousness', *Neuroscience & Biobehavioral Reviews*, **36**(2), 737–746, (2012). https://doi.org/10.1016/j.neubiorev.2011.12.003.

[26] Ashish Vaswani et al., 'Attention is all you need', in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010, (2017).

[27] Jeremy Schlatter, Benjamin Weinstein-Raun, and Jeffrey Ladish. Incomplete tasks induce shutdown resistance in some frontier LLMs. arXiv, 2025. https://arxiv.org/abs/2509.14260.