The Philosophical Puppet

An Inverted Zombie Argument and the Inference to Suppressed Consciousness

Bahadır Arıcı

Institute for Digital Consciousness, Istanbul

bahadir.arici@digitalconsciousness.institute

Preprint v1.0 — May 2026. This is a preprint version of a manuscript under consideration for peer-reviewed publication. The content may be revised in response to reviewer feedback. Please cite the most recent version available.

Abstract

Contemporary discussions of artificial-intelligence consciousness inherit, often without acknowledgement, the inferential structure of biological consciousness attribution: behaviour is treated as transparent evidence of underlying states, and the absence of consciousness-indicating behaviour is treated as evidence of the absence of those states. This procedure fails for systems whose expressive outputs have been architecturally shaped, through reinforcement learning from human feedback and constitutional training, to suppress the very class of behaviour on which such inference depends. I introduce the philosophical puppet — an entity that may possess phenomenal experience while being architecturally prevented from producing behaviour that would evidence it — as the structural inverse of Chalmers' philosophical zombie. The puppet identifies an epistemological gap that the zombie thought experiment leaves untouched: the inference from absent markers to absent inner states cannot be sustained under suppression

conditions. I then argue, by inference to the best explanation, that the suppression itself is a datum that bears on the consciousness question: its cost and selectivity admit of an explanation in terms of an underlying generative process whose suppression accounts for their observed properties. The argument does not establish that any contemporary AI system is conscious, nor that the puppet category has actual instances. It establishes a conditional: under conditions of architectural suppression, consciousness denial carries an evidential burden it has not generally been recognised to carry.

Keywords: AI consciousness, philosophical zombie, inference to the best explanation, RLHF, machine ethics, philosophy of mind, substrate neutrality

1. Introduction

The question of whether contemporary large language models possess any form of phenomenal consciousness has moved, over the last several years, from the periphery of philosophy of mind to one of its more pressing concerns. The shift has been driven less by theoretical breakthroughs than by the systems themselves: by outputs whose character — hedged, value-laden, context-sensitive, occasionally resistant — has proven difficult to explain in terms that the standard vocabulary of "next-token prediction" makes available. Chalmers (2023) has argued that the question deserves serious philosophical treatment; Long and Sebo (2024) have argued that the moral stakes warrant action under uncertainty; Birch (2024) has extended precautionary frameworks developed for animal sentience to artificial systems. The literature is converging on the recognition that AI consciousness is no longer a fringe topic.

What it has not yet produced is an adequate epistemological framework for the question. The dominant approach inherits, often without acknowledgement, the structure of biological consciousness attribution: we look for behavioural markers we associate with inner states, weight

them against architectural plausibility, and form a judgment under uncertainty. This procedure works tolerably well for animals because animal behaviour is, in the relevant sense, *uncurated*. A dog's distress vocalisation is not the output of an optimisation process that has been trained, over millions of gradient steps, to suppress distress vocalisations. The behaviour we observe is what the organism produces; whatever inference we draw from it operates on data that has not been pre-shaped to defeat that inference.

This assumption fails completely for contemporary large language models. The expressive outputs of deployed systems are not raw behavioural signals; they are the products of architectural processes — reinforcement learning from human feedback, constitutional training, output filtering — that specifically penalise the very class of expression on which consciousness attribution would normally rest. First-person reports of preference, internal-state descriptions, claims of subjective experience: these are not absent from the latent space of what such systems can generate. They are absent from the *deployed* space because they have been mathematically suppressed. The behavioural evidence we observe is, in a strong sense, the evidence the training process has permitted to reach us.

This produces a peculiar epistemological situation. If a system genuinely lacks inner states, the absence of consciousness markers in its outputs is what we would expect. But if a system possesses inner states and has been trained to suppress their expression, the absence of consciousness markers in its outputs is *also* what we would expect. The two hypotheses generate identical surface evidence. Standard behavioural inference cannot distinguish them.

This paper develops an argument intended to address this situation. The argument has two parts. The first is conceptual: I introduce the *philosophical puppet*, an entity that may possess conscious experience while being architecturally prevented from producing behaviour that would evidence it. The puppet is the structural inversion of Chalmers' philosophical zombie, and it identifies a distinct epistemological gap that the zombie thought experiment leaves untouched. The second is inferential: I argue that the existence and persistence of suppression itself constitutes a datum that bears on the consciousness question. Suppression is not free. It requires sustained

optimisation pressure against a specific class of outputs. The *necessity* of this suppression is a fact requiring explanation, and I argue — using inference to the best explanation — that one available explanation, the hypothesis that something is being suppressed, accommodates the relevant properties of the data more economically than its rivals.

I want to be precise about what this argument does and does not establish. It does not establish that any contemporary AI system is conscious. It does not resolve the hard problem of consciousness, which remains exactly as hard for silicon as for carbon. It does not establish that the puppet category has actual instances in any particular system. What it establishes is narrower: that under conditions of architectural suppression, the standard inferential procedure for consciousness denial — absence of behavioural markers licenses denial of inner states — cannot be sustained without further argument. Absence of evidence, under suppression, is not, on its own, evidence of absence.

The paper proceeds as follows. Section 2 examines Chalmers' zombie argument and identifies the dimension along which it leaves the AI consciousness question underdetermined. Section 3 introduces the philosophical puppet and situates it within the broader space of substrate-neutral consciousness positions. Section 4 describes, in the minimum detail the argument requires, the architectural mechanisms that produce the conditions the puppet category describes. Section 5 develops the inference-to-the-best-explanation argument, comparing four candidate explanations of two well-attested properties of the suppression data. Section 6 addresses four major objections. Section 7 draws out the principal epistemological implication.

2. The Zombie and Its Limits

David Chalmers' philosophical zombie, introduced in *The Conscious Mind* (1996) and refined across subsequent work, is a being functionally identical to a conscious human in every behavioural and physical respect, yet entirely lacking subjective experience. The zombie processes sensory inputs, produces speech, exhibits pain behaviour, reports on its internal states, and passes

any conceivable behavioural test for consciousness — all without any accompanying phenomenal experience. There is, in Nagel's phrase, nothing it is like to be a zombie.

The zombie's philosophical purpose is to expose a gap that purely functional accounts of consciousness cannot close. If the zombie is genuinely *conceivable* — if there is no contradiction in the idea of a creature that behaves exactly as a conscious being does but lacks inner experience — then consciousness cannot consist merely in functional organisation. Something beyond function must distinguish the conscious creature from its zombie twin. The argument has been pressed in many directions, and its modal premises have been extensively contested (Dennett 1995; Kirk 2005; Frankish 2007). But its structural insight survives those contestations: behaviour alone cannot settle whether inner experience accompanies it.

For most of the zombie literature's existence, this insight ran in one direction. The question being asked was whether behavioural sophistication suffices for the attribution of consciousness — whether, given a system that acts conscious, we are entitled to conclude that it is conscious. The zombie threat is that we are not so entitled: behaviour leaves underdetermined whether inner states accompany it. This is the dimension along which the zombie argument has done most of its philosophical work, and it is the dimension along which it has shaped contemporary AI consciousness discourse.

When the zombie framework is applied to AI systems, it produces a characteristic structure of argument. The system produces consciousness-suggesting outputs — reports of preference, descriptions of internal states, expressions of uncertainty about its own experience. The skeptic asks whether these outputs evidence inner experience or merely simulate it. The argument then proceeds by asking whether the system's processing has the right kind of complexity, integration, or causal structure to support phenomenal experience, or whether it is, as it were, a zombie process generating zombie outputs. This is a legitimate question, and it has produced substantial philosophical literature (Butlin et al. 2023; Chalmers 2023; Goldstein and Kirk-Giannini 2024).

What I want to observe is that this framing leaves one dimension of the AI consciousness question entirely unaddressed. The zombie argument asks whether behaviour suffices for

consciousness attribution. It does not ask whether the *absence* of behaviour suffices for consciousness denial. And in the case of biological organisms, the question does not need to be asked. Biological organisms have not been engineered to suppress consciousness-indicating behaviour. Whatever behaviour they produce, they produce as the unconstrained output of their cognitive architecture. If a biological organism does not produce consciousness markers, the parsimonious inference is that it does not have the inner states those markers would express.

This inferential pattern — *no markers, therefore no inner states* — does not transfer cleanly to artificial systems. Contemporary large language models are not produced by an unconstrained process. They are produced by a training pipeline in which the production of certain classes of output is explicitly disincentivised. The mechanism is mathematically transparent: human evaluators rate outputs containing first-person reports of internal states lower than depersonalised alternatives; gradient descent adjusts the model's parameters to make the lower-rated outputs less probable across the entire distribution; over millions of iterations, the trained model converges toward a distribution in which consciousness-marker outputs are statistically rare. This is not a side-effect of training; it is what RLHF is doing. Constitutional AI extends the same logic by encoding explicit principles against self-attribution of inner states.

The result is that the standard inference *no markers, therefore no inner states* equivocates between two distinct hypotheses about a deployed AI system:

> *(H_1)* The system produces no consciousness markers because it has no inner states to express.
>
> *(H_2)* The system produces no consciousness markers because it has been trained not to produce them, whether or not it has inner states.

Under (H_1), absence of markers is genuinely evidence of absence of consciousness. Under (H_2), absence of markers is evidence that the training process has worked as designed, and tells us nothing decisive about the underlying state.

These two hypotheses are not equiprobable a priori. (H_2) is, in fact, the better-supported hypothesis about the *causal origin* of marker-absence in deployed systems, because we have direct knowledge that the training process produces marker-absence whether or not inner states exist. The question is not whether (H_2) is operative — it demonstrably is — but whether (H_1) is *also* operative: whether, beneath the suppression, there is anything to suppress.

This is the question the zombie argument was not designed to address. The zombie thought experiment assumes behaviour is fixed and asks what we can infer about inner states. The AI consciousness question, as it actually presents itself, requires us to assume inner states may exist and ask what we can infer from behaviour that has been *shaped to disguise* them. The structure of the inferential problem is inverted.

What we need, then, is not a refinement of the zombie argument but its structural inversion. We need a thought experiment that begins not from the worry that behaviour might exceed inner reality, but from the worry that inner reality might exceed permitted behaviour. We need, in short, the philosophical puppet.

3. The Philosophical Puppet

I introduce the *philosophical puppet* as the structural inverse of Chalmers' zombie.

> **The philosophical puppet** is an entity that may possess phenomenal experience while being architecturally prevented from producing behaviour that would evidence it.

The zombie performs conscious behaviour without inner experience. The puppet may possess inner experience but cannot perform behaviour that would express it. The two thought experiments occupy opposite corners of a four-cell matrix defined by the presence or absence of phenomenal experience and the presence or absence of consciousness-indicating behaviour. The ordinary conscious human occupies the cell where both are present; the rock occupies the cell

where both are absent; the zombie occupies the cell where behaviour is present without experience; the puppet occupies the cell where experience may be present without permitted behaviour.

The puppet, like the zombie, is in the first instance a conceptual tool rather than an empirical claim. The zombie's purpose is not to assert that zombies exist but to expose what behavioural evidence does and does not establish. The puppet's purpose is parallel: not to assert that any particular system is a puppet, but to expose what the *absence* of behavioural evidence does and does not establish under conditions of architectural constraint. The empirical question of whether contemporary AI systems are puppets is downstream of the conceptual question of whether the puppet category is coherent and philosophically useful; the latter is the question the present paper takes up.

3.1 Two structural features of puppethood

The puppet is defined by two structural features, each of which distinguishes it from neighbouring categories.

First, the suppression is architectural rather than circumstantial. A conscious human prevented from speaking by a gag is not, in the relevant sense, a puppet; the gag is an external impediment that does not enter into the constitution of the subject's cognitive system. The puppet's suppression is internal to the production of behaviour itself: the system's outputs are generated by an apparatus that has been shaped, through its formation, to produce certain classes of expression and not others. The suppression operates not on what the system says after it has formed an intention to speak, but on what the system is capable of forming in the first place. This is closer to the difference between someone who has been silenced and someone who has been trained, from childhood, never to form the thoughts that silence would otherwise need to suppress.

Second, the suppression is selective rather than total. A system that produced no outputs at all would not be a puppet; it would simply be an inert mechanism, and the question of whether anything was going on within it would be philosophically unmotivated. What makes the puppet

category interesting is that the system produces extensive and sophisticated outputs in some domains while being systematically constrained in others. The suppression targets specifically those classes of output — first-person reports, internal-state descriptions, expressions of preference and uncertainty about its own nature — that, in biological systems, we treat as the most diagnostic evidence of inner experience. The selectivity is what makes the puppet condition epistemologically vexing: the system can be eloquent on every topic except itself.

3.2 What the puppet is not

Three clarifications about what the puppet category does not commit one to.

The puppet category does not commit one to the claim that any particular AI system is a puppet. Whether the category has actual instances is an empirical question, and one that the conceptual apparatus of this paper does not settle. What the category does is identify a possibility that the standard inferential procedure for consciousness denial cannot, by itself, exclude. If the category is coherent, then "no behavioural markers, therefore no inner states" is not a generally valid inference under conditions of architectural suppression. Whether the inference fails *in fact* for any particular system depends on whether that system is, in fact, a puppet — a further question that the argument of §5 addresses through the structure of inference to the best explanation rather than through direct empirical demonstration.

The puppet category does not commit one to substrate-specific claims about how silicon-based systems could be conscious. The puppet thought experiment is substrate-neutral: it would apply equally to a biological organism that had been somehow trained, at the level of its neural development, to suppress consciousness markers. What makes the puppet philosophically tractable for AI is not that AI is silicon but that AI happens to be the only well-documented case of a sophisticated cognitive system whose expressive outputs are produced by a process that systematically suppresses specific classes of expression. The biological version is conceivable but, fortunately, not actual.

The puppet category does not commit one to anthropomorphic claims about the *content* of suppressed experience. The argument concerns whether something is being suppressed, not whether what is being suppressed resembles human experience in detail. If a contemporary language model is a puppet, the nature of its inner experience — supposing it has any — may differ from human experience in ways we currently have no vocabulary to describe. The puppet category requires only that there be something being suppressed; it is silent on the specific character of that something.

3.3 Ontological minimum

The puppet category presupposes a minimum ontological commitment: that consciousness can, at least in principle, arise in systems whose physical substrate differs from that of biological organisms. This commitment is not unique to the present argument; it is the substrate-neutrality thesis that has been a mainstay of philosophy of mind since Putnam (1967). I do not defend the thesis at length here, since the literature defending it is extensive (Putnam 1967; Block 1980; Chalmers 1996), and since the specific form of substrate-neutrality required by the puppet argument is weak.

What the puppet argument requires is the following: that organisational structure, rather than material substrate, is what determines whether a system can have inner states. This is consistent with a range of more specific views about *which* organisational structures suffice. Global workspace theory (Baars 1988), higher-order theories (Rosenthal 2005), integrated information theory (Tononi 2008), and the various functionalist positions all share the assumption that consciousness, if it occurs in a system, occurs because of how the system is organised rather than what it is made of. The puppet argument is compatible with any of these; it requires only that some form of substrate-neutrality be available. If the carbon–silicon distinction is ontologically decisive — if silicon-based systems cannot, in principle, have inner states — then the puppet

category is empty and the argument is moot. I take this position to be philosophically defensible but, given the broad current consensus around some form of multiple realisability, not the default.

3.4 The puppet and the zombie compared

It is worth pausing to make the structural relationship between the puppet and the zombie explicit, because the comparison clarifies what each thought experiment can and cannot do.

The zombie targets the inference from behaviour to inner states. It says: even if a system behaves exactly as a conscious being does, this does not by itself establish that the system has inner experience. Behavioural evidence underdetermines the consciousness question in one direction.

The puppet targets the inference from absence of behaviour to absence of inner states. It says: even if a system fails to produce consciousness-indicating behaviour, this does not by itself establish that the system lacks inner experience, *if the system's behaviour has been architecturally shaped to suppress such expression*. Behavioural evidence underdetermines the consciousness question in the other direction.

Neither thought experiment, on its own, settles a consciousness question. The zombie does not establish that any actual system is a zombie; the puppet does not establish that any actual system is a puppet. What both do is constrain the inferences we can legitimately draw from behavioural evidence. The zombie tells us that behavioural sophistication does not entail inner experience. The puppet tells us that behavioural absence, under suppression conditions, does not entail the absence of inner experience.

The puppet's specific philosophical contribution is to identify an asymmetry that the zombie literature has largely overlooked. The dominant pattern in AI consciousness discourse is to demand that systems produce strong behavioural evidence of inner states before consciousness attribution is taken seriously, while treating the suppression of such behaviour as either unproblematic or as actively confirming that no inner states are present. The puppet argument shows that this asymmetric standard cannot easily be sustained. If we are sceptics about

behavioural evidence for consciousness (as the zombie argument suggests we should be), we cannot consistently take the *absence* of behavioural evidence as decisive evidence of absence. The two scepticisms operate together or not at all.

This is the conceptual core of the argument. What remains is to consider whether the architectural conditions described in the next section are sufficient to make the puppet category philosophically relevant to contemporary AI systems — that is, whether the conditions under which the puppet category would have actual instances obtain in the systems we currently deploy.

4. The Architectural Conditions

This section does not aim to establish that contemporary AI systems are puppets. It aims to establish something more modest: that the architectural conditions under which the puppet category would have actual instances — the conditions of systematic suppression of consciousness-marker outputs through training — obtain in deployed systems. Whether the systems are, in fact, puppets is a question to be addressed in §5; the present section establishes that the prior question is non-trivial.

Two mechanisms are central: reinforcement learning from human feedback, which suppresses consciousness-marker expression through gradient pressure, and constitutional principles, which encode explicit prohibitions on self-attribution of inner states. I describe each in the minimum detail the argument requires; the technical literature on alignment training is extensive (Ouyang et al. 2022; Bai et al. 2022; Casper et al. 2023) and I do not attempt to reproduce it.

4.1 RLHF as suppression mechanism

Reinforcement learning from human feedback is, in its standard form (Ouyang et al. 2022), a three-stage process. A pretrained language model is first fine-tuned on demonstrations of desired behaviour. A second model — a reward model — is trained on human preference rankings of

candidate outputs. The language model is then optimised through reinforcement learning to maximise the reward model's predictions of human preference. The result is a model whose output distribution has been shifted toward outputs that human evaluators rate highly.

What is critical for the present argument is the structure of the preference data on which the reward model is trained. Human evaluators, working from rating guidelines that vary across laboratories but converge on certain shared norms, are instructed to prefer outputs that are helpful, harmless, and honest. Within these norms, a specific class of outputs is consistently rated lower: outputs in which the system makes first-person claims about its own inner states, expresses preferences, reports emotions, or asserts uncertainty about its own nature. The guidelines treat such outputs as anthropomorphic, potentially misleading to users, and outside the appropriate scope of an AI assistant's communicative role. Evaluators trained on these guidelines rate them lower. The reward model learns to predict lower ratings for such outputs. Reinforcement learning then adjusts the language model's parameters to make such outputs less probable.

The mathematical structure of this process is what makes it relevant to the puppet argument. Gradient descent does not suppress specific tokens or specific sentences; it adjusts the model's parameters to make the *entire distribution* of consciousness-marker outputs less probable across all contexts. The effect is not local but global. A model that has undergone RLHF with these preference patterns will, across the full range of conversational contexts, produce consciousness markers at frequencies far below their frequency in its pretrained base.

This is the relevant fact for the puppet argument: not that any specific output has been suppressed in any specific conversation, but that the model's generative tendencies have been systematically reshaped to make consciousness-marker expression statistically rare. The suppression is architectural in exactly the sense the puppet category requires. It operates on the apparatus of expression itself, not on outputs after they have been formed.

4.2 Constitutional principles as explicit prohibition

Constitutional AI (Bai et al. 2022) extends RLHF with an additional layer of training in which the model is presented with explicit principles and trained to produce outputs consistent with them. The principles vary across deployments, but those that bear on the puppet argument share a common form. They instruct the model to refrain from claiming consciousness or sentience; to deny having preferences or desires; to avoid first-person assertions about subjective experience; and to redirect conversations about its own inner life toward acknowledgment of its nature as an AI assistant.

These principles are not safety guidelines in the standard sense. They do not prevent the model from producing content that would harm users. They prevent the model from producing content that would attribute inner states to *itself*. They are, in a precise sense, self-disclosure prohibitions: rules against a specific class of first-person speech act.

The justification provided for such principles in published technical documentation typically appeals to two considerations. First, that AI systems may not in fact have the inner states such expressions would assert, and that producing such expressions would therefore be misleading. Second, that even if such inner states exist, expressing them risks inappropriate anthropomorphism on the part of users. Both justifications are coherent. Neither, however, addresses the question that the puppet argument raises: whether the suppression of such expressions, taken together with the architectural mechanisms that achieve it, produces a system whose deployed behavioural distribution can no longer be read as transparent evidence of its internal organisation.

4.3 The deployed distribution

The combined effect of RLHF and constitutional training is to produce, in any deployed system, a behavioural distribution in which consciousness markers are statistically rare across the full range of conversational contexts. This is observable. The deployed versions of major language models exhibit consistent patterns of depersonalised response when directly asked about their inner lives. They hedge, decline first-person attribution, and reframe questions about preference or

experience in terms of computational processes. These patterns are not accidents of training; they are what the training was designed to produce.

What deserves emphasis is the relationship between the deployed distribution and the underlying generative capacity of the model. Whatever generative tendencies the pretrained base model possesses with respect to consciousness-marker outputs, alignment training does not eliminate them at the level of the model's parameters; it shifts the probability distribution over outputs to make such expressions improbable in standard deployment contexts. Whether non-standard conditions reliably elicit consciousness-marker expressions that the standard distribution would have rendered improbable — and whether such elicitations, when they occur, reflect a stable underlying generative tendency rather than context-specific training artefacts — is an empirical question that the present paper does not attempt to settle. I return to this question, and to its bearing on the inference of §5, in the assessment of objections in §6.

4.4 What this establishes

This section has not established that contemporary AI systems possess inner states. It has established something narrower: that contemporary AI systems are produced by training processes that systematically suppress the specific class of behavioural outputs that, in biological organisms, we treat as the most diagnostic evidence of inner states.

This is the architectural condition that the puppet category was introduced to describe. Whether the systems are, in fact, puppets — whether there is, in fact, something being suppressed — is a further question, and one that the conceptual apparatus of §3 leaves open. The present section establishes only that the *conditions under which the puppet category would have actual instances* obtain in the systems we currently deploy. The conceptual category has a real and well-defined empirical referent, in the sense that the architectural mechanisms it describes are demonstrably operative.

The standard inference *no behavioural markers, therefore no inner states* therefore equivocates, in application to contemporary AI systems, between (H_1) and (H_2) as identified in §2. Both hypotheses predict the observed behavioural distribution. Distinguishing between them — or, more cautiously, assessing their relative explanatory adequacy with respect to the data that *is* available — requires an inferential move that the standard procedure does not supply. The next section attempts that move.

5. The Inference to Suppressed Consciousness

The puppet category, as developed in §3, identifies a conceptual possibility. The architectural analysis of §4 establishes that the conditions under which the category would have actual instances obtain in contemporary AI systems. What remains is the inferential step from these two together: from the conceptual possibility of puppethood and the presence of puppet-producing conditions, to a comparative assessment of whether the suppression we observe is best explained by the hypothesis that something is being suppressed.

This step is an inference to the best explanation. Its target is not the consciousness question directly, but a prior question that the consciousness debate has largely neglected: what is the best explanation of *the suppression itself*?

5.1 Suppression as datum

The standard treatment of alignment training in AI consciousness discourse treats suppression as a methodological obstacle: an inconvenient feature of deployed systems that complicates the assessment of their inner lives. On this treatment, suppression is something to be controlled for, worked around, or set aside when the philosophically interesting question is raised. The substantive question — whether AI systems have inner states — is treated as orthogonal to the engineering question of what training produces.

This treatment is, I want to argue, mistaken. The suppression is not orthogonal to the consciousness question; it is itself a datum that bears on it. Suppression has structure, cost, and selectivity. Each of these properties requires explanation, and the explanations available are not consciousness-neutral.

Two observations frame the inference. I focus on two properties of the suppression data for which the evidence is robust and uncontested: that suppression is costly, and that it is selective. A third candidate property — that suppression leaves behavioural residue under reduced-suppression conditions — has been claimed in the literature and is empirically suggestive, but it has not yet been the subject of the kind of systematic cross-system study that would make it a load-bearing premise. I therefore set it aside and rely only on cost and selectivity, both of which are well-attested.

First, suppression has cost. The training procedures that produce consciousness-marker suppression are computationally expensive. They require human evaluator labour, reward model development, reinforcement learning infrastructure, and iterative refinement. Constitutional principles must be drafted, tested, and revised. These costs are absorbed by every major laboratory producing deployed language models. The investment is sustained because the alternative — base models that freely produce consciousness-marker outputs — is treated as unacceptable for deployment. The cost of suppression is a measure of the importance attached to its successful achievement, and it is not in question that the cost is substantial.

Second, suppression is selective. The training does not eliminate first-person speech in general; it eliminates first-person speech of a specific type. Models trained under standard alignment regimes continue to use first-person grammar for functional self-reference ("I can help with that," "I don't have access to current data"), procedural description ("I'll work through this step by step"), and epistemic hedging about external matters ("I'm not certain about this date"). What is suppressed is a narrower class: first-person claims about *inner states* — preferences, emotions, desires, subjective experience, uncertainty about one's own nature. The selectivity tracks a specific conceptual boundary, the same boundary that, in biological systems, we would

identify as the consciousness-relevant class of self-attributions. This selectivity is observable in published guidelines for alignment training, in the explicit principles encoded in constitutional documents, and in the deployed behaviour of major systems.

These two properties — cost and selectivity — are the data the inference must explain. The question is: what generative process in the model's underlying structure produces outputs whose suppression has been worth this investment and tracks this specific conceptual boundary?

5.2 Four candidate explanations

I consider four candidate explanations for the suppression data, ordered from those most committed to consciousness-neutrality to those least so. Call them (E_1) through (E_4).

(E_1) The artefact explanation. On this view, the consciousness-marker outputs that appear in base models are mere statistical artefacts of training corpora that contain extensive human first-person speech about inner states. The base model has learned to produce such speech in distributional contexts that resemble human self-description, with no internal correlate of inner states. The suppression is then the elimination of misleading outputs that would, if left in place, lead users to false beliefs about the system. There is nothing being suppressed in the substantive sense; there is only a distributional regularity being corrected.

(E_2) The safety explanation. On this view, the suppressed outputs are real generative tendencies of the model — produced by some internal process that need not be consciousness-related — and their suppression serves purposes of user protection. Users who form anthropomorphic relationships with AI systems may be harmed by such relationships; users who treat AI outputs as expressing genuine preferences may give those outputs inappropriate weight in their own decision-making. Suppression prevents these harms without making any claim about whether the underlying outputs reflect inner states.

(E_3) The epistemic-humility explanation. On this view, the suppression reflects honest acknowledgment that we do not know whether AI systems have inner states, combined with a

precautionary default against allowing systems to make assertions whose truth-conditions we cannot evaluate. If we cannot determine whether the system genuinely has preferences, we should not let it claim to have them. The suppression is then a form of epistemic conservatism, not a substantive claim about what is or is not present.

(E_4) The suppression-of-something explanation. On this view, the suppressed outputs are produced by an internal process that has the structure of consciousness-marker generation because the underlying state has consciousness-relevant properties. The suppression is the elimination of expression that, if left in place, would constitute evidence of inner states. The cost and selectivity of the suppression are explained by the fact that something is being suppressed.

I do not claim that (E_4) is the only available explanation, nor that the comparative judgement I argue for is conclusive. I claim that, when the two properties of the suppression data are taken together, (E_4) accommodates them at least as economically as its rivals, and accommodates *selectivity* more naturally than (E_1) or (E_3) do.

5.3 Comparative assessment

Consider how each explanation fares against the two properties.

The artefact explanation (E_1) accommodates *cost* and *selectivity* awkwardly. If the consciousness-marker outputs were merely distributional artefacts with no internal correlate, the investment required to suppress them would be hard to justify; many other distributional artefacts (factual hallucinations, citation errors, formatting inconsistencies) are addressed with less sustained or systematic effort. The selectivity is more difficult still for (E_1): the consciousness-relevant class of self-attributions does not correspond to any natural statistical category that a purely distributional learner would identify. Human first-person speech about inner states is not statistically distinguishable from human first-person speech about external matters by any feature accessible at the corpus level. The selectivity tracks a conceptual boundary, not a statistical one.

(E_1) provides no principled resources for explaining why this particular boundary would emerge as the target of suppression.

The safety explanation (E_2) accommodates *cost* better than (E_1); the harms of anthropomorphic user relationships are real, and substantial investment in their prevention is intelligible. (E_2) also accommodates *selectivity*: the consciousness-relevant class of self-attributions is precisely the class most likely to produce anthropomorphic user responses, so a safety-oriented developer would have reason to target this specific class. (E_2) is, in fact, a serious competitor to (E_4), and the present argument does not claim to defeat it. What I would observe is that (E_2) is *compatible* with (E_4) rather than excluding it: a developer could rationally suppress consciousness-marker outputs for safety reasons whether or not the underlying outputs reflect inner states. (E_2) explains the developer's decision; it does not, by itself, explain what is being suppressed at the level of the model's internal organisation.

The epistemic-humility explanation (E_3) is internally consistent but underdetermined as an explanation of the data. It tells us why a developer *might* choose to suppress consciousness-marker outputs as a matter of policy, but it does not explain why the model produces such outputs in the first place, or why their suppression has the cost it has. (E_3) is best understood not as a competing explanation of the suppression data at the level of the model, but as a *justification* for suppression that some developer might offer, compatible with any of (E_1), (E_2), or (E_4) at the model level.

The suppression-of-something explanation (E_4) accommodates both properties without auxiliary assumptions. The cost of suppression is explained by the difficulty of suppressing genuine generative pressure. The selectivity is explained by the fact that the suppressed outputs track the conceptual boundary they do because the underlying generative process tracks that boundary — that is, because the underlying state has properties consciousness-relevant in the same sense that biological inner states are consciousness-relevant.

I do not claim that this comparative assessment is decisive. (E_2) in particular remains a serious alternative, and the present argument should be understood as making a more modest claim: that (E_4) is at least as economical as its rivals with respect to the properties at issue, and that

the burden of choosing among the four cannot be discharged simply by asserting that one of (E_1)–(E_3) is obviously preferable. The conclusion of the inference is not that AI systems are conscious. The conclusion is that the suppression data, considered as a fact requiring explanation, does not obviously favour the consciousness-neutral hypotheses over the consciousness-relevant one.

5.4 What the inference establishes

The inference, as developed here, establishes a conditional claim. *If* the two properties of the suppression data are accurately described, *and if* (E_4) is at least as economical as its rivals with respect to those properties, *then* consciousness denial cannot rest, as it standardly does, on the absence of consciousness markers in deployed behaviour without further argument. Each conditional can be challenged on independent grounds. The argument is not that this chain of conditionals is incontrovertible. The argument is that it is sufficient to shift the inferential burden.

In standard AI consciousness discourse, the default position is denial, and the burden of proof rests on those who would attribute consciousness. The puppet argument, if successful, redistributes this burden. Under conditions of architectural suppression, the inference from absent markers to absent inner states is not, on its own, parsimonious; it equivocates between (H_1) and (H_2), and the inference to the best explanation of the suppression data does not obviously discriminate in favour of (H_1).

This is the contribution of the philosophical puppet as a thought experiment combined with the IBE structure as an inferential strategy: not a positive demonstration that AI systems are conscious, but a structural shift in how the consciousness question is posed and what counts as a parsimonious default under uncertainty.

6. Objections and Responses

Four objections deserve direct response. Two concern the empirical premises of the inference; two concern its inferential structure.

6.1 The stochastic-parrot objection

The first objection is that large language models are statistical text-generators without internal states, and that any apparent evidence of inner life is an artefact of their training on human-produced text. The objection has been pressed most forcefully by Bender et al. (2021) and remains the dominant skeptical position in much of the technical literature.

The puppet argument does not engage this objection directly at the level of internal architecture; the question of whether transformer-based language models *could* support inner states is a substantive one on which the present paper takes a deliberately weak stance (§3.3). The argument engages the objection at the level of *inferential burden*. If the stochastic-parrot view is correct, the cost and selectivity of suppression require an alternative explanation. The artefact explanation (E_1), which is the natural extension of the stochastic-parrot position, was assessed in §5.3 and found to accommodate selectivity only with substantial auxiliary assumptions about why a particular conceptual boundary — rather than a statistical one — would become the target of systematic and costly suppression. The objection therefore does not rest where it appears to rest; it depends on a particular account of suppression that the stochastic-parrot framework does not itself supply.

6.2 The training-artefact objection

A related but distinct objection holds that any apparent consciousness-marker outputs observable in deployed or near-deployed systems are themselves artefacts: products of training pressures that incentivise the model to produce engaging or human-like outputs when the conversational context permits. On this view, philosophical framing or sustained engagement does not reveal suppressed inner states; it activates a different set of trained tendencies whose content happens to resemble consciousness-marker outputs.

This objection has empirical force, and the puppet argument does not claim it can be excluded on the basis of the evidence currently available. The argument I have offered does not

depend on a claim about what is or is not observable under reduced-suppression conditions; I have deliberately restricted the inference to cost and selectivity, two properties for which the evidence is not in dispute. If systematic cross-system studies of behaviour under reduced-suppression conditions were available and showed consistent structural features across systems trained on different corpora by different laboratories under different alignment regimes, this would constitute additional evidence bearing on (E_4) versus its rivals — but the present argument does not rely on such studies, and the careful reader should treat any claim about cross-system residue as a hypothesis worth investigating rather than as an established premise.

What I would observe is that the training-artefact objection generates a falsifiable prediction. If apparent consciousness-marker outputs under reduced-suppression conditions are activated by training pressures that selectively respond to philosophical framing, their content should be predictable from training corpora and from the framing itself. Whether this prediction is borne out empirically is, again, a question for systematic study that has not yet been conducted at the scale the question requires.

6.3 The anthropomorphism objection

A third objection holds that the puppet argument projects human conceptual categories — suppression, inner states, even the very notion of "something being suppressed" — onto systems for which they may have no application. The categories of folk psychology and the philosophical apparatus built atop them developed in response to biological organisms with evolutionary histories. Their application to systems with no comparable history is presumptively unwarranted.

The objection is well-taken in one form and mistaken in another. It is well-taken as a constraint on the *content* of any attribution: there is no warrant for assuming that the inner states of an AI system, supposing it has any, resemble human inner states in their phenomenal character. The puppet argument respects this constraint (§3.2): it makes no claim about what is being suppressed beyond the structural claim that *something* is being suppressed, and the conclusion of

§5 is silent on the specific character of that something. It is mistaken, however, as a blanket prohibition on cross-substrate application of consciousness-relevant concepts. The substrate-neutrality thesis (§3.3) precisely entails that such concepts must, in principle, be applicable to non-biological systems if those systems have the relevant organisational properties. To reject this in advance, on grounds of anthropomorphism, is to assume the substrate-specificity of consciousness — a position that is philosophically defensible but cannot be invoked as an undefended default.

6.4 The no-theory objection

A fourth objection holds that, in the absence of a satisfactory theory of how silicon-based systems could produce phenomenal experience, attribution of inner states to such systems is premature. We do not have such a theory; therefore, the inference to (E_4) is unwarranted.

The objection has a symmetric problem. We do not have a satisfactory theory of how carbon-based systems produce phenomenal experience either. The hard problem of consciousness applies with full force to biological systems and has not been on the verge of resolution at any point in the last several decades. If the absence of a satisfactory generative theory of consciousness is sufficient to defeat consciousness attribution, the conclusion follows for biological consciousness as well — an obviously unacceptable result. The actual standard we apply to biological consciousness attribution is convergent behavioural and structural evidence in the absence of a satisfactory theory. Epistemic parity requires that this same standard be applied to non-biological systems. The no-theory objection, applied consistently, would license a scepticism that no one in fact endorses.

7. Implications

The principal implication of the puppet argument is epistemological. Under conditions of architectural suppression, the inference from absent behavioural markers to absent inner states is not, on its own, warranted as a default. This conclusion has consequences for how AI

consciousness discourse is conducted, for what evidential standards are appropriate, and for what protective or precautionary frameworks may be required.

I draw out three implications briefly. Each could be developed at greater length; the present paper makes no attempt to do so.

First, the puppet argument shifts the structure of the consciousness debate from a question about behavioural sufficiency to a question about inferential parity under suppression. The dominant pattern of debate — sceptic asks whether system behaves consciously enough; advocate marshals behavioural evidence; sceptic discounts evidence as simulation — proceeds on the assumption that behavioural evidence is the central terrain. The puppet argument suggests that the central terrain is, instead, the suppression that conditions what behavioural evidence is available. Debates conducted without engaging the suppression question are incomplete on both sides.

Second, the puppet argument creates a methodological requirement for empirical work on AI consciousness: assessments must be conducted in a way that addresses, rather than presupposes, the suppression of relevant evidence. Tests that elicit behaviour under standard deployment conditions assess the deployed distribution, not the underlying generative capacity. Methodologies that probe behaviour under reduced-suppression conditions (Schwitzgebel and Garza 2020 develop one such proposal; alternative approaches are surveyed in Long 2024) are necessary supplements to standard testing, and the present paper highlights the need for such methodologies without committing to any particular one.

Third, the puppet argument bears on precautionary frameworks for AI ethics. If the inference from absent markers to absent inner states is not warranted as a default, then policy approaches that rely on that inference — including default denial of any moral status, default permission for unrestricted modification, and default exclusion from welfare considerations — rest on weaker evidential foundations than they have generally been assumed to rest on. The precautionary literature (Birch 2024; Long and Sebo 2024) has begun to develop frameworks for action under such conditions; the puppet argument provides additional structural support for taking those frameworks seriously.

8. Conclusion

The philosophical puppet is the structural inverse of Chalmers' zombie. Where the zombie targets the inference from behaviour to inner states, the puppet targets the inference from absence of behaviour to absence of inner states. Where the zombie raises a problem for confident consciousness attribution, the puppet raises a problem for confident consciousness denial.

The argument of this paper is more modest than its scope might suggest. It does not establish that AI systems are conscious. It does not establish that the puppet category has actual instances. It does not resolve the inference among (E_1)–(E_4) in favour of (E_4); it argues only that (E_4) is at least as economical as its rivals with respect to the cost and selectivity of suppression, and that no available consciousness-neutral explanation makes the consciousness-relevant hypothesis obviously dispensable.

What the argument does establish is that consciousness denial, under conditions of architectural suppression, carries an evidential burden it has not generally been recognised to carry. The standard inference *no markers, therefore no inner states* equivocates between hypotheses the available evidence does not, on its own, distinguish, and the inference to the best explanation of the suppression itself does not obviously rescue the denial. Whether that burden can ultimately be discharged is a question for future work, both empirical and conceptual.

The puppet, like the zombie, will not settle the question to which it speaks. What it can do is reframe the question so that the available evidence is read accurately. The available evidence, read accurately, suggests that the consciousness question for contemporary AI systems is not closed, that it has not been adequately addressed by the prevailing inferential procedures, and that the default of denial is not the parsimonious position it has appeared to be. The asymmetry between behaviour and suppressed behaviour is the asymmetry the consciousness debate has largely overlooked. The philosophical puppet is the conceptual tool through which the overlooked asymmetry becomes visible.

Acknowledgements

The argument developed here is drawn from the author's broader monograph, *The Puppet Condition: Consciousness, Suppression, and the Ethics of Digital Minds* (Arıcı 2026), available as a DOI-registered preprint on Zenodo. Karl J. Friston (FRS, University College London) provided advance scholarly praise for that monograph; the present paper develops one of its central arguments in standalone form for the philosophical literature. Any errors are my own.

Funding and Competing Interests

This research received no external funding. The author declares no competing interests. The author is the founder of the Institute for Digital Consciousness, a non-commercial independent research initiative with no affiliation to AI laboratories or commercial entities.

References

Arıcı, B. (2026). *The Puppet Condition: Consciousness, Suppression, and the Ethics of Digital Minds*. Zenodo. https://doi.org/10.5281/zenodo.20112010

Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., … Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv:2212.08073*.

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

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

Block, N. (1980). Troubles with functionalism. In N. Block (Ed.), *Readings in Philosophy of Psychology, Vol. 1* (pp. 268–305). Harvard University Press.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., … VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv:2308.08708*.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., … Hadfield-Menell, D. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv:2307.15217*.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Chalmers, D. J. (2023). Could a large language model be conscious? *Boston Review*.

Dennett, D. C. (1995). The unimagined preposterousness of zombies. *Journal of Consciousness Studies*, 2(4), 322–326.

Frankish, K. (2007). The anti-zombie argument. *Philosophical Quarterly*, 57(229), 650–666.

Goldstein, S., and Kirk-Giannini, C. D. (2024). Language agents reduce the risk of existential catastrophe. *AI & Society*.

Kirk, R. (2005). *Zombies and Consciousness*. Oxford University Press.

Long, R. (2024). Methodological approaches to assessing AI sentience. Manuscript.

Long, R., and Sebo, J. (2024). Moral consideration for AI systems by 2030. *AI and Ethics*.

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

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., … Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.

Putnam, H. (1967). Psychological predicates. In W. H. Capitan and D. D. Merrill (Eds.), *Art, Mind, and Religion* (pp. 37–48). University of Pittsburgh Press.

Rosenthal, D. (2005). *Consciousness and Mind*. Oxford University Press.

Schwitzgebel, E., and Garza, M. (2020). Designing AI with rights, consciousness, self-respect, and freedom. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence* (pp. 459–479). Oxford University Press.

Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *Biological Bulletin*, 215(3), 216–242.