The Solipsist Wall

A Two-Sided No-Go Argument for Theories of Consciousness

antra

2026.11.22 Anima Labs Field Note FN-2026-005

Version `f9dc2f3d` of 2026-11-23. Current version: https://animalabs.ai/field-notes/solipsist_wall/

Abstract

Can evidence settle which systems are conscious? We argue that it cannot: the attempt runs into a formal boundary, the *Solipsist Wall*. Combining Kleiner and Hoel's third-person falsification dilemma with Caplain and Reason's first-person no-go argument, we show that under structural closure, phenomenal realism, and first-person certainty, wherever the bridge rule from structure to experience is not fixed in advance, no experiment can establish it, and no internal detector can be certified beyond cases already accepted. Evidence can show which features track experience in those cases and how closely another system resembles them, but not how much resemblance is enough: rival criteria that agree on those cases fit all observations equally well, and simplicity priors cannot break the tie from data. We map the main responses and their costs. For AI consciousness, the wall points to the futility of waiting for empirical evidence to settle the dispute: where to draw the line is an extra-empirical commitment that should be argued for openly.

These are working notes, and they are still in progress.

Most debates about consciousness—especially now, as people argue over whether language models have inner lives—proceed as if we are one clever experiment or one better interpretability tool away from settling the question. Physicalists look for biological or causal-structure correlates; functionalists look for representational statefulness, global broadcast, or computational valence; skeptics demand proof before extending moral concern.

There is a structural reason this debate never resolves empirically. It isn't a temporary shortage of neuroscience or mechanistic interpretability. It is a formal boundary—what we can call the **Solipsist Wall**.

Two independent lines of formal work expose this boundary from opposite sides:

- **The third-person falsification dilemma** (Doerig et al., 2019; Kleiner & Hoel, 2021), which shows that the standard way of testing theories of consciousness is caught between two horns: if reports are trusted wherever they occur, theories that predict experience from internal structure are falsified by behaviorally identical substitutes, while theories that define experience by report or behavior are unfalsifiable. Kleiner and Hoel name two ways out: a "lenient" dependency between internal prediction and behavioral inference that is neither of these (they know of no existing example), or a physics that is not causally closed.

- **The first-person certainty no-go argument** (Caplain, 1995; developed by Reason, 2016, 2019), which argues that no physical or computational process can be *certain* that it is conscious, because no such process can validate its own reliability from inside.

Notably, the original authors draw very different conclusions from these results: Reason argues for a violation of physical causal closure (specifically energy conservation), while Hoel's later work (2025) deploys the falsification dilemma to argue *against* LLM consciousness. Here, we synthesize both sides into a single formal framework—not to force a single metaphysical answer, but to map the exact boundary of what empirical science can settle and to set out the main **responses** available to a theory of consciousness.

1. The Setup

Let:

- $\mathcal{S}$ be the space of **realizable structural states and processes**—physical, biological, or computational. Everything measurable about a system's internal constitution (neural activity, fMRI scans, weights, KV caches, activations, causal graphs) is part of or computed from $s \in \mathcal{S}$.
- $\mathcal{O}_{\text{beh}}$ be the space of **behavioral and report observables** (verbal reports, text outputs, button presses, external actions), and $\pi_{\text{beh}} : \mathcal{S} \to \mathcal{O}_{\text{beh}}$ the map giving each state's behavior and reports.
- $\mathcal{E}$ be the space of **phenomenal states** (first-person subjective experiences), containing a distinguished null element $\emptyset \in \mathcal{E}$ representing the absence of experience.
- A **Bridge Rule** (or Theory of Consciousness) be a map $B : \mathcal{S} \to \mathcal{E}$, with associated conscious set $\mathcal{P}_B := \{s \in \mathcal{S} : B(s) \neq \emptyset\}$.
- An **Inference Map** be a rule $I : \mathcal{O}_{\text{beh}} \to \mathcal{E}$ used by third-person experimenters to infer experience from behavior and reports.
- A **structural model** M be a complete account of the structural side: which states are realizable, how they evolve, and how probable each observation is under each experiment. The lemmas below hold for any given M; §5 takes up what changes when M itself is uncertain.

Following Kleiner and Hoel (2021), notice the clean separation of roles: internal measurements (brain scans, probes, circuit topology) sit on the **prediction side** ($\mathcal{S} \xrightarrow{B} \mathcal{E}$), while external reports and behaviors sit on the **inference side** ($\mathcal{O}_{\text{beh}} \xrightarrow{I} \mathcal{E}$). Empirical testing compares the two: does $B(s) = I(\pi_{\text{beh}}(s))$?

2. Minimal Axioms

- **A1 (Structural Closure):** All structural transitions in $\mathcal{S}$ and all behavioral observables depend solely on $\mathcal{S}$, independently of B. Formally, π_{beh} does not take B as an argument: replacing B with any other bridge rule B' leaves every internal state $s \in \mathcal{S}$ and every behavioral observable $\pi_{\text{beh}}(s)$ identical. In terms of probabilities: given the structural model, every third-person observation is exactly as likely under one bridge rule as under another, in every experiment, including interventions and experiments chosen in the light of earlier results.

- **A2 (Phenomenal Realism):** Experience is real, distinct from its absence, and instantiated by at least one realizable system: $|\mathcal{E}| \geq 2$ and $\mathcal{P}_B$ is non-empty.
- **A3 (First-Person Certainty / Unspoofable Acquaintance):** You can know directly, without risk of being fooled, that you are having an experience right now—even if you could be mistaken about everything else. Formally, there is a first-person epistemic state K ("direct certainty that experience is occurring") that is **sound** over all realizable states: no state $s \in \mathcal{S}$ instantiates K while $B(s) = \emptyset$. The observer whose evidence is at issue is in a state s_0 that instantiates K; this is the first-person datum. It says that experience is present at s_0, and nothing about whether K holds in any other state. The Claim in §5 needs only this datum; that K is sound is used in §4 and in Responses 2, 3 and 5.
- **A4 (Behavioral Substitution / Unfolding):** Internal structure makes finer distinctions than behavior does. An internal property $\mathcal{P} \subsetneq \mathcal{S}$ (such as biological substrate, recurrent network topology, or high integrated information Φ) has a **behavioral twin** when two realizable systems behave identically on the outside—$\pi_{\text{beh}}(s) = \pi_{\text{beh}}(s')$—even though one has the property ($s \in \mathcal{P}$) and the other lacks it ($s' \notin \mathcal{P}$). **A4** is the assumption that a given property has one. It is made property by property and cannot hold for all of them: a property defined by behavior alone has no twin. *(Doerig et al., 2019, give one example: a finite-horizon recurrent system can be unfolded into a behaviorally equivalent feedforward network. O'Reilly-Shah, Selvitella & Schurger, 2026, show that no single static feedforward network can do so over arbitrary time intervals when the recurrent network's weights change rapidly.)*
- **A5 (The Bridge Is Open):** At the state t in question, neither answer—conscious or not—is ruled out before the evidence comes in: among the bridge rules that agree on the **accepted cases** (the states, including the observer's own, where experience is taken as given) and have not been excluded in advance, some count t as conscious and some do not. When this holds, the bridge is **open at** t. It is a statement about what has been ruled out, not about what is possible.

A1 carries most of the argument, so it is worth saying what it does and does not claim. It says that what a system does next, and what it reports, is fixed by its structural state, up to whatever chance the structural laws themselves contain. Experience is not an additional input to the dynamics. Three things it does not say:

- It does not say that today's physics is complete. $\mathcal{S}$ is all structure, including physics not yet discovered. A new force or field found at work in brains can be described as new structure, and whether **A1** holds is then asked of the enlarged $\mathcal{S}$.
- It does not say that experience is causally inert. On an identity view experience has effects, because it *is* a structural property.
- It is not determinism. The structural laws may be probabilistic; **A1** says only that the probabilities do not depend on B.

Three definitions are used from here on:

- **Admissible hypothesis:** Roughly, an admissible bridge rule is any rule consistent with your own experience at s_0. Formally, which states instantiate K is not a structural fact given in advance; it is part of what a hypothesis says. A **hypothesis** is therefore a pair (B, K_B): a bridge rule, and the set of states in which certainty holds under it. It is

admissible for an observer if it satisfies **A2** and **A3**, that is, if $s_0 \in K_B \subseteq \mathcal{P}_B$. Since K_B can always be taken to be $\{s_0\}$, a bridge rule belongs to some admissible hypothesis exactly when it assigns experience to s_0, and such a rule is called admissible too.

- **Accepted cases:** Any set of states, containing the observer's own, on which the assignment of experience is taken as given. Suppose you discovered that your own brain contained no neurons. Your certainty that you are having an experience would give you a counterexample to any rule that requires neurons. *(An observer passes through many states, and everything below holds with the set of them in place of s_0. First-person evidence enters only as the fact that experience is present in those states. Treating oneself as a random sample of all the conscious states there are would let the number of other candidate systems bear on the question; that is a disputed way to reason and is not assumed here.)*
- **Readout and spoofability:** Call $D : \mathcal{S} \to \{0, 1\}$ a **readout** if its value is set by a mechanism distinct from the property it reports on—a belief register, a self-monitoring circuit, a report channel. A readout is **spoofable** if there exists a realizable spoof state $s_{\text{spoof}} \in \mathcal{S}$ (via lesion, noise, or causal interception) such that $D(s_{\text{spoof}}) = 1$ while $B(s_{\text{spoof}}) = \emptyset$. Membership in $\mathcal{P}_B$ is a property of the state, not a readout of it. *(Showing that a readout can come apart from its usual cause is not enough for spoofability: the spoof state must have no experience at all, which is itself a claim about B. Caplain's "cognitive separation" principle is weaker: it allows a recording method that only ever records truths, and says only that this reliability cannot be checked from inside.)*

On **A5**: with no prior restriction on bridge rules, the bridge is open at every state outside the accepted cases. A prior philosophical commitment can close it—everywhere, or only in certain cases. An identity claim, on which experience simply *is* a specific structural property, closes the bridge everywhere because it excludes every rule but one. A narrower principle like organizational invariance (Chalmers, 1996, ch. 7)—which holds that any two systems with the same fine-grained functional organization have the same experience—closes the bridge only for exact functional replicas of accepted cases (such as a neuron-for-neuron silicon copy of a human brain), while leaving it open for systems organized differently, like a fish or a language model. Notice that structural closure (**A1**) is separate from whether the bridge is open (**A5**): an identity theorist accepts **A1** and closes the bridge by prior commitment instead. Why closing the bridge by prior commitment does not solve the empirical problem is taken up under Response 3.

3. Wall One: Third-Person Uncertifiability (Extending Kleiner–Hoel)

Lemma 1 (Third-Person Uncertifiability): *Under **A1** and **A2**, third-person observations (internal or behavioral) can test a bridge rule $B : \mathcal{S} \to \mathcal{E}$ only against an assumed inference map I, and can certify neither. Moreover, if I is taken as valid across all of $\mathcal{S}$:*

1. *Any B that assigns different experiences to two behaviorally identical states is falsified by I; every theory built on an internal property for which **A4** holds is of this kind.*
2. *Any B defined as a function of behavior is either already falsified by I or agrees with it by construction; measurements of internal structure add no information.*
3. *A B built on an internal property that has no behavioral twin can pass a genuine test against I, but passing certifies neither B nor I.*

Proof. Suppose an experimenter tests B using an inference map $I : \mathcal{O}_{\text{beh}} \to \mathcal{E}$, assumed valid on a calibration domain $\mathcal{S}_{\text{cal}} \subseteq \mathcal{S}$ (e.g., awake verbal humans).

1. **Outside the calibration domain ($\mathcal{S}_{\text{cal}} \subsetneq \mathcal{S}$):** Define an alternative bridge rule B' that agrees with B on $\mathcal{S}_{\text{cal}}$ and differs outside it—mapping every state in $\mathcal{S} \setminus \mathcal{S}_{\text{cal}}$ to $\emptyset$, say, or permuting the experiences assigned there. By **A1**, every internal state $s \in \mathcal{S}$ and every behavioral observable $\pi_{\text{beh}}(s)$ is identical under B and B'. Hence B is empirically underdetermined outside $\mathcal{S}_{\text{cal}}$, wherever such a B' has not been excluded in advance (**A5**).
2. **Universal inference domain ($\mathcal{S}_{\text{cal}} = \mathcal{S}$):** Either B separates two behaviorally identical states or it does not, and if it does not, that holds either by the theory's definition or as a matter of fact. These three cases line up roughly with the three regimes Kleiner and Hoel distinguish between prediction data and inference data (independence, strict dependence, and lenient dependency):
   - **A behavioral twin exists** (Kleiner and Hoel's substitution argument, which applies whenever prediction and inference are independent). Suppose there are states s, s' with $\pi_{\text{beh}}(s) = \pi_{\text{beh}}(s')$ and $B(s) \neq B(s')$. Since $I(\pi_{\text{beh}}(s)) = I(\pi_{\text{beh}}(s'))$, I disagrees with B at s or at s'. The two cannot both be right everywhere: B **is falsified under** I. Every B whose conscious set is an internal property for which **A4** holds is in this case: take $s \in \mathcal{P}_B$ and $s' \notin \mathcal{P}_B$. The unfolding argument is the special case of a recurrent network and its feedforward twin.
   - **Strict dependence** (B is a function of behavior by definition). Suppose the theory defines $B = f \circ \pi_{\text{beh}}$ for a stated $f : \mathcal{O}_{\text{beh}} \to \mathcal{E}$, as in behaviorism or input–output functionalism (Kleiner and Hoel suggest that global workspace theory arguably belongs here too). Whether f agrees with I is settled by comparing the two rules on the behaviors that occur; no measurement of internal structure bears on it. The theory is either already falsified ($f \neq I$ on some behavior) or unfalsifiable ($f = I$)—in Kleiner and Hoel's words, "empirically uninformative or tautological".
   - **No behavioral twin, as a matter of fact** (one special case of lenient dependency). Suppose B is built on an internal property $\mathcal{P}$ for which **A4** fails: no realizable state outside $\mathcal{P}$ behaves identically to one inside it. Which behaviors accompany $\mathcal{P}$ must then be found by experiment, so agreement between B and I is a genuine test, neither failed in advance nor empty. Kleiner and Hoel (2021) leave lenient dependency open and knew of no theory that had it. O'Reilly-Shah, Selvitella and Schurger (2026) prove that no single static feedforward network can reproduce, over arbitrary time intervals, a recurrent one whose weights change rapidly, and offer plasticity as a concrete instance; Hoel (2025) argues that theories grounded in continual learning are another.
3. **In every case the test is relative to I.** Even when a theory passes every test against I, we can modify B and I together—say, declaring a class of non-human behaviors unconscious, or relabeling experiences—without changing any observable. Formally, suppose (B, I) passes every test. Pick any behavior o, produced by no accepted case, to which I assigns an experience, and let B' and I' agree with B and I except that they assign $\emptyset$ there: $I'(o) = \emptyset$, and $B'(s) = \emptyset$ whenever $\pi_{\text{beh}}(s) = o$. The pair (B', I') passes every test as well and, by **A1**, no observable differs, yet the two pairs disagree about whether those states are conscious. The same holds for contents: relabel the experiences by any

one-to-one $\sigma : \mathcal{E} \to \mathcal{E}$ that fixes $\emptyset$ and every experience assigned in an accepted case, and $(\sigma \circ B, \sigma \circ I)$ reproduces every agreement and every disagreement of (B, I).

Thus third-person empirical testing can check consistency between a theory B and an assumed inference rule I, but can certify neither. □

4. Wall Two: First-Person Closure (after Caplain and Reason)

The natural move around Wall One is to anchor B in the one place where experience is known directly: **first-person certainty**. If a physical or computational mechanism inside $\mathcal{S}$ accounts for how I know with certainty that I am conscious, perhaps we can extrapolate from that mechanism.

Wall Two shows that no such mechanism can be certified as a detector of experience beyond the cases already accepted, from outside or from inside.

> **Lemma 2 (First-Person Closure):** *Assume **A1**. If a readout $D : \mathcal{S} \to \{0, 1\}$ fires at a state where the bridge is open (**A5**), then it is unsound under some bridge rule that agrees on the accepted cases and has not been excluded in advance. So no readout can be certified sound at such a state, even if it in fact never errs. With no restriction on bridge rules, that covers every state outside the accepted cases.*

Proof. Call a readout sound under a bridge rule if it never fires in a state to which that rule assigns no experience. By **A1** the readout gives the same output in each state whichever bridge rule is true. If D fires at a state t where the bridge is open, some rule that agrees on the accepted cases and has not been excluded counts t as not conscious, and under that rule D fires without experience. Conversely, a readout that fires only at states which every such rule counts as conscious is sound under all of them, so those states are exactly how far certification reaches. With no restriction on bridge rules the bridge is open at every state outside the accepted cases: take any admissible rule that agrees on them and change its value at t. □

Two consequences follow.

- **Checking does not help.** A second readout that checks the first, a formal verification of it, or any tower of self-checks is more structure, and its outputs too are the same under every bridge rule. This is the point of Caplain's cognitive separation principle and of Reason's regress: a record cannot validate its own reliability from inside, even if it never errs.
- **A readout that can be spoofed is not certainty.** If D is a spoofable readout, some realizable state has $D = 1$ and no experience. If D constituted K, that state would instantiate K without experience, which **A3** forbids.

Neither consequence says that K is non-physical. Together they say that if certainty is a readout, it is one that can never be spoofed, and that nothing could certify this of a readout that fires where the bridge is open. Under **A1**–**A3**, that leaves two possibilities, taken up as Responses 3 and 5 below: K is not a readout at all, or it is a readout that never errs and cannot be shown to be one.

Caplain (1995) and Reason (2016) argue for more than this lemma: that a physical system

cannot have the certainty that conscious subjects have, where certainty means something the subject can validly guarantee to itself. The lemma takes from them only the point about certification, and **A3** asks only for a state with no false positives.

The Key Distinction: Inferential Readout (D) vs. Direct Acquaintance (K)

A sharp critic will push back on **A3**: *"You've smuggled the conclusion into K. Of course no zombie has felt certainty, because 'felt' presupposes experience; meanwhile, my actual belief 'I am conscious' is just a fallible structural readout $D(s) = 1$ that could in principle misfire."*

That objection is exactly right—and it clarifies the real fork. Ask the critic: **Can you know *with certainty* right now that you are not in s_{spoof} (a zombie whose belief-circuit D misfired)?**

- If **yes**—because you are directly acquainted with experience rather than inferring it by reading a fallible wire—then **A3** holds, and by the second consequence of Lemma 2 that certainty is not a spoofable readout. You are in **Response 3** or **Response 5** below (or **Response 4**, if you also give up closure).
- If **no**—because your belief that you are conscious really is only a reliable-but-fallible readout—then you have rejected **A3** and taken **Response 2** below.

5. The Combined Boundary and the Main Responses

The two lemmas are two faces of one boundary, the **Solipsist Wall**. In plain terms:

> Evidence can show which features accompany experience in the cases we already accept, and how closely another system resembles those cases. It cannot show how much resemblance is enough. For any way the world might be built, any two criteria that agree on the accepted cases fit all third-person evidence equally well, now and in the future, however much they disagree about animals or machines. Learning about structure can change a verdict on a particular system, but only by applying a rule one already holds for getting from structure to a verdict. That rule is never tested, and complete knowledge of how the world is built would leave it exactly where it was.

The Stack

A system can resemble us at several levels. From the bottom up: what it is made of, its biology, the architecture of its processing, the computation it carries out, and what it does and says. Call these levels the **stack**. Three questions can be asked about any proposed criterion for consciousness, and evidence bears on only two of them:

1. **Which feature tracks experience in the accepted cases?** Evidence can settle this, relative to trusting reports. It is what experiments on human subjects do.
2. **Where on the stack does a given system land?** How closely an animal or a model resembles us, level by level, can be measured.
3. **Where does the cutoff go?** How much resemblance is enough, and at which level, is not settled by evidence, however complete: every answer fits the evidence equally well.

The reason for the third answer is that in humans every level of the stack comes together. "Global broadcast", "global broadcast in neurons", and "global broadcast in a primate brain" fit every human experiment equally well. They disagree only about systems in which the levels come apart, and those are exactly the systems for which we have no independent access to experience. A language model resembles us near the top of the stack and departs at the bottom; a fish does roughly the reverse. So the choice is less how far resemblance must reach than which level counts. This is close to what Block (2002) called the harder problem of consciousness.

This is stronger than saying the evidence is not yet in. Evidence does change verdicts on particular systems, in three ways. In each, the work is done by a link that is assumed and never tested:

- *A case already accepted rules a criterion out.* People with atypical brains or with implants, who report as others do, show that some features are not required. That works only because their reports are already trusted, and which cases to trust is the cutoff question itself. The same holds of one's own case: an observer who learned that her own substrate contained no neurons could drop a neurons-only criterion.
- *A verdict changes when we learn more about a system.* Learning that an animal or a model has some feature changes the verdict on it only for someone who already holds that the feature is what matters. That is applying a cutoff, not revising one.
- *A theory wins in the human case.* If the winning theory is taken to come with a level, its success seems to carry over to machines, or to count against them. The packaging is an assumption: the same theory restricted to neurons fits the human evidence exactly as well.

The pattern can be made exact. Suppose two structural models, M_1 and M_2, are equally likely beforehand, and two investigators agree on what each model predicts. They differ only in the rule that takes a structural world to a verdict on some target system:

	Conscious if M_1	Conscious if M_2
Investigator A	90%	10%
Investigator B	10%	90%

Both start at 50%. An experiment then establishes M_1. A moves to 90% and B to 10%. They made the same predictions, started from the same odds, saw the same result, and moved in opposite directions. What moved them was the rule each brought, and no experiment tests that rule: by **A1** every such rule gives the same probability to every observation. Evidence tells us which structural world we are in. It does not tell us how to get from a structural world to a verdict, and complete knowledge of structure would leave that question exactly as open.

There is one exception, and it cannot be shared. A system that is itself conscious has first-person evidence about its own state. If a model were conscious, it would have evidence about the cutoff that no observation could give anyone else, and it could not pass that evidence on to anyone who does not already trust its reports: by **A1** those reports are the same whichever bridge rule is true.

The Precise Statement

Fix a structural model M and a set of accepted cases, and let t be a state outside them. Admissible rules are as defined in §2.

> **Claim (The Solipsist Wall):** *Assume **A1–A3**, and that the bridge is open at t (**A5**). Then:*
>
> 1. ***No experiment discriminates.*** *Any two admissible bridge rules that agree on the accepted cases give the same probability to every third-person observation, in every experiment, and pass or fail the same tests on the accepted cases. Given M, the odds between them are the same after any observation as before.*
> 2. ***Both answers stay open.*** *Two such rules, neither excluded in advance, disagree about t: on one, t is conscious; on the other, it is not.*
> 3. ***No test is reliable under both.*** *Given M, a procedure that declares t conscious or not on third-person evidence gives each answer with the same probability under those two rules, and an answer that is right under one is wrong under the other. Its chances of error under the two add up to one.*
> 4. ***Learning structure does not learn the rule.*** *If M itself is uncertain, evidence changes the probability of each structural model and leaves the probability of each bridge rule, given the model, exactly where it was. A verdict on t can then move only through the link a person assumes between structural models and bridge rules, and, for a given view of structure, every such link gives the same probability to every observation.*
>
> *So no third-person evidence, however complete, settles whether t is conscious, and none tests the rule that would settle it. First-person evidence fixes only the observer's own states.*

The first two parts are immediate. By **A1** no observation depends on the bridge rule, in any experiment, and two rules that agree on the accepted cases pass or fail the same tests there, so given M the likelihood ratio between them is exactly 1 for all data, present and future. The second part is **A5** at t. With no restriction on bridge rules it holds at every state outside the accepted cases: take any admissible rule that agrees on them and change its value at t. The third part follows from the first two: the same evidence is equally likely under both rules, so a procedure gives each verdict equally often under both, and the two rules make opposite verdicts correct. The fourth part is Bayes' rule. The probability of an observation, given a structural model and a bridge rule, does not depend on the bridge rule, so it cancels, and the probability of each bridge rule given the model is unchanged. This is how Bayesian updating treats any quantity that the data do not identify (Poirier, 1998). Two people who share a view of structure and differ only in the link they assume predict every observation identically, so no observation can decide between their links. When M is uncertain, new observations can only shift your odds between two bridge rules if you already link them to different expectations about structure—as Investigators A and B did above. If your expectations about structure do not depend on which bridge rule is true, your odds between them never move.

The Claim uses neither **A4** nor the two lemmas. Lemma 1 shows how the wall appears inside the standard way of testing theories. Lemma 2 shows why the one case that is not assumed cannot be extended by finding a detector. The observer's own case is a state, not an individual: any system in exactly the state s_0 gets the same verdict, because B is a function of the state.

Every other accepted case is accepted by trusting an inference map there (Lemma 1), and no readout found in another system can be certified to carry that trust further (Lemma 2).

None of this says that all bridge rules are equally plausible. There can be reasons to prefer one, and some reasons are better than others: Chalmers's fading-qualia argument (1996, ch. 7), for example, is a reason to accept organizational invariance. On the stack, that principle is a cutoff at the level of functional organization: it says the levels below do not matter, and it is silent on how much less than exact duplication is enough. The claim is that such reasons are not observations, and that no observation will stand in for them.

The Main Responses

Rather than declaring a single metaphysical winner, the argument sorts the main responses, each with a clear theoretical cost:

1. **Reject A2 — Eliminativism / Strong Illusionism** *(e.g., Dennett, Frankish)*
   Deny that there are phenomenal states in the sense assumed here: there are structural states, internal models, and reports that represent experience as having such a character, and $\mathcal{P}_B$ is empty. *(This is not a denial that people have experiences in the everyday sense.)*
   - *Cost:* Denies the reality of the first-person datum most people treat as their starting point.
2. **Reject A3 — First-Person Fallibilism / Reliabilism**
   Maintain that experience is real, but drop infallible certainty: hold that first-person awareness of being conscious is a reliable-but-fallible structural readout D. *(Reason's argument explicitly sets aside confidence short of full certainty.)*
   - *Cost:* You concede that you cannot rule out with certainty that you are a zombie right now—and you remain fully inside **Lemma 1**, since the reliability of D cannot be empirically calibrated against $\mathcal{E}$ without presupposing an unverified inference rule I.
3. **Keep A1–A3 — Certainty Is Not a Readout, and the Bridge Is Primitive**
   Accept both lemmas. First-person certainty K is real and sound, and it is not a detector wired to $\mathcal{P}_B$. There are two readings of what it is instead:
   - **Constitutive identity** *(e.g., Brute Identity Theory, Axiomatic Functionalism)*: experience simply *is* a specific structural property $\mathcal{P}_B$, and K is being in it. This reading closes the bridge everywhere (**A5**): it excludes every rule but one before any evidence comes in. K is then structural but not a readout, so the question of spoofing does not arise.
   - **Lawful accompaniment** *(e.g., property dualism with closed physics, Russellian Monism, Panprotopsychism)*: experience is something further that the structure brings with it or rests on.
   - *Cost, on both readings:* Knowing that you are conscious still leaves open which property of your present state makes that true. Third-person measurements can identify many properties you have in that moment. Each candidate leaves a further question open. Why does having that property feel like anything at all? On either reading, choosing B supplies the **extra-empirical axiom** that settles which other states carry experience.

- *Further cost, on the second reading:* By **A1**, every report of certainty—this note's assertion of **A3** included—would have been produced identically under any other bridge rule (Chalmers's "paradox of phenomenal judgment", 1996).

4. **Reject A1 — Non-Closed Physics / Interactionist Dualism** *(Reason's own conclusion; one of the two ways out Kleiner & Hoel name)*
   Assert that experience makes a lawful difference to what a system does: hold the structural state fixed, change the bridge rule, and the system would evolve differently (for example, under a law where the next state depends on both the current structural state and the experience it carries). On such a law, **A1** fails. Kleiner and Ludwig (2024) call the condition they need *dynamical relevance*: whenever a system is conscious, it evolves differently from what an underlying physical theory says. They take it to be weaker than a violation of causal closure, and they argue that empirical tests of theories of consciousness presuppose it, since consciousness must make a difference to the data. That is close to the contrapositive of our main claim: if the bridge makes no difference to the data, the data cannot test it.
   - *What it gains:* Given the law by which experience acts, rival bridge rules predict different observations, so experiments can tell them apart, and reports of experience are caused by experience. This is the one response here that predicts something new to observe. Chalmers and McQueen (2022) work out one such proposal, on which consciousness collapses the quantum wave function: its simplest versions are already ruled out, and others could in principle be tested with quantum computers.
   - *What it does not gain:* An escape from the Solipsist Wall. Any **lawful** causal role can be redescribed as structure: enlarge $\mathcal{S}$ to a space $\mathcal{S}'$ with a new variable that plays exactly the role the law gave to experience. Any experiment confirming the new dynamics equally confirms that purely structural description. **A1** then holds of $\mathcal{S}'$, and wherever the bridge is still open (**A5**) the Solipsist Wall re-forms around it: is that new physical variable actually felt? Discovering a new causal contribution singles out a new physical phenomenon, not a phenomenal interpretation; closing that gap still requires a prior commitment. Conversely, an influence that followed *no* law at all—not even a statistical one—would resist absorption into $\mathcal{S}'$, but nothing about it could be tested or studied scientifically.
   - *Cost:* Even to single out such a new physical candidate (short of escaping the wall), this view requires a consciousness-specific deviation from standard physical dynamics with no established experimental evidence, plus a law saying which matter carries it—either a sharp line somewhere in evolution and in development where the dynamics change, or an effect present in all matter. For current AI it cuts one way: processors are designed and verified to follow the specified computation and to suppress deviations, so if consciousness is dynamically relevant in their sense, systems running on them are not conscious (Kleiner & Ludwig, 2024).

5. **Keep A1–A3 — Certainty Is a Readout That Cannot Be Spoofed**
   Hold that there is a mechanism D, distinct from $\mathcal{P}_B$, that never fires without experience, and that first-person certainty is that readout. Caplain's principle allows this: a recording method may record only truths.
   - *Cost:* D would be a consciousness meter, valid wherever it fires, but where the bridge is open nothing can certify which readout it is or that it never errs—not third-person tests (**Lemma 1**), and not the readout itself or any check on it (**Lemma 2**).

*(Note: these responses are not claimed to be exhaustive. Rejecting **A4** for some internal property—one route to lenient dependency, as in Hoel (2025)—avoids falsification but not the last step of Lemma 1: I itself remains uncertified.)*

6. Why Solomonoff Induction Cannot Settle It from Data

A common attempt to rescue **Response 2** or **Response 3** without admitting an extra-empirical choice is to invoke **Minimum Description Length (MDL) / Solomonoff induction**: *"Even if we can't verify B directly, we can favor the bridge rule with the shortest description—for example, a biological/physicalist law, a causal-structure law like IIT, or a computational/functionalist law like stateful representational control or global workspace broadcast."*

Where the data go on favoring one hypothesis over another, so that its likelihood advantage grows without bound, the choice of reference machine washes out: it shifts description lengths by at most a constant, and an unbounded advantage outweighs any constant (Rathmanner & Hutter, 2011). That guarantee is about predictions. It was never a promise that two hypotheses with the same predictions would be told apart.

Here the predictions are the same. Under **A1**, two rival bridge rules B_{phys} and B_{comp} attached to the same structural model predict **identical distributions over all third-person data** (both internal measurements of $\mathcal{S}$ and behavioral observables $\mathcal{O}_{\text{beh}}$). Their likelihood ratio is identically 1, so **observation never changes the odds between them**: the posterior odds equal the prior odds, whatever the prior was. A fixed reference machine does rank the two rules, so simplicity can break the tie. But the ranking is not forced by data, and it is not canonical. Take any universal machine used to describe complete hypotheses, structure and bridge together, and write the two hypotheses down as two distinct finite strings. Follow the machine with the map that swaps those two strings in its output and leaves every other output as it is. The result is another universal machine that makes exactly the same predictions about every observation and ranks the two rules the other way round. Predictive success cannot say which machine was the right one to use.

This is narrower than it may sound. A theory of consciousness usually carries a structural commitment as well—a mechanism, a signature in neural dynamics—and that part is testable in the ordinary way: the preregistered adversarial tests of integrated information theory and global neuronal workspace theory (Cogitate Consortium, 2025) challenged key predictions of both. Evidence of this kind answers the first question on the stack: which features track experience in the accepted cases. It leaves the third untouched: how far beyond those cases a feature may be carried.

7. What Follows for Science and AI

This argument does not render consciousness research useless—it draws a sharp line between what empirical science *can* resolve and what requires an explicit **extra-empirical** commitment. Of the three questions asked on the stack, two are empirical.

1. What tracks experience in accepted cases. Experiment can settle this, given that we trust reports: rival theories make testable predictions about neural and cognitive signatures in human subjects, and those predictions can fail.

2. Which internal features a given system actually has. Measurement can settle this—everything inside $\mathcal{S}$ is open to empirical investigation:

- We can empirically test whether a system maintains stateful internal representations of its own condition (via computational complexity bounds, perturbation studies, and mechanistic interpretability).
- We can measure whether a system integrates evaluation criteria along a unified valence axis, tracks prediction error, models itself and others, or conceals internal states from its surface outputs.
- We can map exact structural isomorphisms and divergences between biological brains and artificial architectures.

3. Which features are sufficient for consciousness (where to draw the cutoff). Neither experiment nor measurement can settle this. Findings of both kinds can change a verdict on a particular system, but only through an assumed, untested rule linking structure to experience. What empirical science cannot produce is a consciousness meter that crosses the Solipsist Wall and returns an unconditional verdict on "true phenomenality." Take two rules that agree on the accepted cases and disagree about some system. Any procedure that examines evidence and declares that system conscious or not gives the same answers whichever of the two rules is true. So it cannot be reliable under both: if it is right nine times in ten when one rule holds, it is wrong nine times in ten when the other does. Finding new physics at work in conscious systems would not change this: it would supply a new physical candidate for the bridge, not an escape from the wall (see Response 4).

Crucially, the formal boundary itself does **not** automatically entail either inflationism or deflationism about AI consciousness—as illustrated by the fact that Hoel (2025) uses the third-person dilemma to argue *against* LLM consciousness, while others use the exact same epistemic limit to argue for precautionary inclusion. What the wall *does* eliminate is the pretense that either stance is forced by third-person data alone:

- Requiring resemblance all the way down to biology, and so treating biological similarity as mandatory for moral patienthood, is an **extra-empirical axiomatic choice**, not a scientific finding.
- Requiring it only at the level of computation or representation (such as stateful self-modeling and integrated valence), and taking that as the practical anchor for consciousness attribution, is *also* an **extra-empirical choice**—one that can and should be argued for openly on rational grounds: theoretical coherence, ethical prudence, and game-theoretic stability under permanent uncertainty.

No external evidence—now or in the future—can test which of these requirements is objectively correct. That choice is empirically underdetermined and has to be argued for in the open, rather than treated as a verdict waiting on the next experiment.

References

- Block, N. (2002). "The harder problem of consciousness." *The Journal of Philosophy*, 99(8), 391–425.
- Caplain, G. (1995). "Is consciousness a computational property?" *Informatica*, 19, 615–619.

- Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press. (Chapter 5, "The Paradox of Phenomenal Judgment", and Chapter 7, "Absent Qualia, Fading Qualia, Dancing Qualia".)
- Chalmers, D. J., & McQueen, K. J. (2022). "Consciousness and the collapse of the wave function." In S. Gao (ed.), *Consciousness and Quantum Mechanics*. Oxford University Press. arXiv:2105.02314.
- Cogitate Consortium (2025). "Adversarial testing of global neuronal workspace and integrated information theories of consciousness." *Nature*, 642, 133–142.
- Doerig, A., Schurger, A., Hess, K., & Herzog, M. H. (2019). "The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness." *Consciousness and Cognition*, 72, 49–59.
- Hoel, E. (2025). "A Disproof of Large Language Model Consciousness: The Necessity of Continual Learning for Consciousness." arXiv:2512.12802 (v3, January 2026).
- Kleiner, J., & Hoel, E. (2021). "Falsification and consciousness." *Neuroscience of Consciousness*, 2021(1), niab001.
- Kleiner, J., & Ludwig, T. (2024). "The case for neurons: a no-go theorem for consciousness on a chip." *Neuroscience of Consciousness*, 2024. arXiv:2304.05077.
- O'Reilly-Shah, V. N., Selvitella, A. M., & Schurger, A. (2026). "A caveat regarding the unfolding argument: implications of plasticity." *Neuroscience of Consciousness*, 2026(1), niag027.
- Poirier, D. J. (1998). "Revising beliefs in nonidentified models." *Econometric Theory*, 14(4), 483–509.
- Rathmanner, S., & Hutter, M. (2011). "A philosophical treatise of universal induction." *Entropy*, 13(6), 1076–1136.
- Reason, C. M. (2016). "Consciousness is not a physically provable property." *The Journal of Mind and Behavior*, 37(1), 31–46. arXiv:1705.00096.
- Reason, C. M. (2019). "A No-Go Theorem for the Mind–Body Problem: An Informal Proof that No Purely Physical System Can Exhibit All the Properties of Human Consciousness." *The Journal of Mind and Behavior*, 40(2), 95–120.