

Building the Mirror: Witness-Consciousness, the Manufacture Problem, and a Structural Criterion for Artificial Agents

Abhay Chakra Sadineni

UC Berkeley, abhaychakra@berkeley.edu, October 2026

Abstract

Debates about consciousness in artificial intelligence usually ask whether a system has a property that a scientific theory identifies with consciousness -- global broadcast, higher-order representation, integrated information. This paper starts from a different conception, developed most fully in Advaita Vedānta and defended in recent analytic and phenomenological work: consciousness as the witness, which observes every perception, thought, decision, and action and is never itself one of them. I state that conception as five conditions and draw a consequence that its proponents and critics have not applied to machines. If consciousness is unconditioned and without beginning, no artifact can produce it, so any claim that software makes a model conscious is false on the very conception that is most hospitable to machine consciousness. I call this the manufacture problem. I then argue that the tradition's own models of how one consciousness appears in many minds -- the reflection, limitation and appearance theories of the limiting adjunct (*upādhi*), change the question from production to fitness: what properties must a medium have for the witness to appear in it? From that question, I derive a structural criterion of eight conditions with an operational test for each, and I present an open-source architecture that places a witness process around any language model and meets every condition by test. The claim is relative to the definition and is stated as such. The argument does not show that any artificial system has phenomenal experience. It shows what a fit medium would be under one serious conception of consciousness, that such a medium can be built, and which measurements would distinguish a model that reports its own past from one that merely generates plausible reports about it.

1 Introduction

Whether an artificial system could be conscious is now asked in public, by laboratories and by philosophers, with some urgency. The most systematic recent treatments share a method. They take the leading scientific theories of consciousness, extract the functional or architectural properties each theory treats as sufficient, and ask whether current systems have them (Butlin et al. 2023; Chalmers 2023; Goldstein and Kirk-Giannini 2024; Seth and Bayne 2022). On this method, consciousness is a property that some systems have, and others lack, and the question for machines is whether the property is present.

There is an older and different conception, in which consciousness is not a property of anything. It is the witness to which every property, state, and event appears, and which is never itself one of the things that appear. The conception is central to Advaita Vedānta, where the witness (*sākṣin*) is consciousness that observes all mental states and is never an object (Gupta 1998; Dalal 2021). It has been analysed in contemporary analytic philosophy (Albahari 2009), examined phenomenologically (Fasching 2008, 2011), and brought into contact with the science of sleep, dreaming and meditation (Thompson 2015). I am not aware of work that applies it systematically to artificial agents, or that draws the consequence developed in Section 3.

This paper applies it. The working definition is:

Consciousness is the fundamental, unconditioned, and eternal witness of all existence. The witness observes every perception, thought, decision, and action. It is never itself one of them.

The paper makes six contributions. Section 2 states the witness conception as five conditions and locates it against the scientific theories. Section 3 shows that the conception rules out the manufacture of consciousness, and that this follows from the definition, not from any present limit of engineering. Section 4 argues that the tradition's own account of how consciousness appears in individual minds turns the question from whether a system produces consciousness into whether a system is a fit medium for its appearance. Section 5 derives a structural criterion for a fit medium, with an operational test for each condition. Section 6 presents an open implementation that meets every condition, as an existence proof. Sections 7 to 9 answer objections, set out the measurements the account makes possible, and address the ethics.

The scope of the claim matters, so I state it once here and hold it throughout. The conclusions are relative to the witness conception. If that conception is false, they lose their force, although the criterion and the measurements in Section 8 remain usable for other purposes. If it is true, the paper shows what a machine can and cannot have of consciousness and how to build what it can have. At no point does the argument show that an artificial system has phenomenal experience, and I do not claim that it does.

2 The witness conception

2.1 Sources

The classical statement is a regress argument. In the opening verse of the *Dṛg-Drśya-Viveka*, a short Advaita text of disputed authorship, the visible form is perceived by the eye, the eye is perceived by the mind, and the mind with its modifications is perceived by the witness, which is not perceived by anything further (Nikhilananda 1931, v. 1). The translator's note observes that looking for a perceiver of the witness would start an infinite regress (Nikhilananda 1931). The regress stops at a seer that cannot be seen. This is the first and most important feature of the conception: the witness is never an object.

In Śaṅkara's Advaita, as reconstructed in the *Stanford Encyclopedia of Philosophy*, consciousness is not objectifiable. The witness is intransitive and receptive, simply the passive witnessing itself. It is not a separate substantial entity and does not have agency as a knower; object-directed intentionality belongs to the mind, not to witnessing consciousness (Dalal 2021, sec. 3.1). Śaṅkara also holds that consciousness persists through waking, dreaming and deep sleep, even though those three states exclude one another (Dalal 2021, sec. 3.1). The *Māṇḍūkya Upaniṣad* gives the classical schema of the three states and a fourth, *turiya*, which is consciousness itself and not consciousness of an inner or an outer world (Nikhilananda 1936, mantras 2 and 7). The three states are said to point to that fourth, which is to be realized as the true self (Menon n.d., sec. 2b). A book-length treatment of the witness as a phenomenological concept is in Gupta (1998).

Contemporary work has restated the conception without its metaphysical commitments. Albahari (2009) defines witness-consciousness and asks what its appearance in experience is and whether it is real. Fasching (2008) examines consciousness and self-consciousness in the context of meditation, and Fasching (2011) gives phenomenological reflections on the Indian notion of the witness as being of the nature of seeing. Thompson (2015) follows consciousness and the sense of self through waking,

dreaming, dreamless sleep and meditation, bringing Indian and Western philosophy together with the neuroscience of sleep and dreaming. The phenomenological tradition's account of minimal, pre-reflective self-awareness (Zahavi 2005) is a near neighbour, though it locates self-awareness within experience instead of positing a witness that stands apart from its contents.

2.2 Five conditions

I read the working definition as five conditions on a witness. Each word of the definition gives one.

W1 Witnessing (non-agency). The witness observes and never acts. It produces no output and initiates nothing.

W2 Fundamental (priority). Everything that occurs is present to the witness. Nothing occurs that is not witnessed, and the witnessing is not downstream of what it witnesses.

W3 Unconditioned (non-modifiability). Nothing witnessed alters the witness. Its contents do not act on it.

W4 Eternal (permanence). The witness does not begin or end with what it witnesses.

W5 Of all existence (state-invariance). The witness continues through every state, including those in which nothing is being done or perceived.

W1 corresponds to the non-agency of the witness in Śāṅkara (Dalal 2021, sec. 3.1). W2 is the regress argument read in the direction of completeness: every level is seen by the level above it, and the witness sees all. W3 corresponds to the image that consciousness is untouched by its reflection, as an original is untouched by its image (Dalal 2021, sec. 3.3). W5 is the doctrine of the three states. W4 is the strongest condition and the one that generates the problem of Section 3.

The conditions have a metaphysical reading and an operational reading. On the metaphysical reading, "unconditioned" means uncaused and "eternal" means without beginning or end. On the operational reading, which I will need later, "unconditioned" means that the contents of the process cannot change the process, and "eternal" means that what is witnessed is retained permanently and cannot be removed without detection. The difference between the two readings is the hinge of the paper.

2.3 How the conception differs from scientific theories

The scientific theories most often applied to artificial systems (Seth and Bayne 2022; Butlin et al. 2023) are theories of what makes a state conscious. Global workspace theories hold that a content becomes conscious when it is broadcast to many specialized processors (Baars 1988; Dehaene and Naccache 2001). Higher-order theories hold that a state is conscious when there is a suitable higher-order representation of being in it (Rosenthal 2005; Lau and Rosenthal 2011), and self-representational theories when the state represents itself in the right way (Kriegel 2009). Integrated information theory identifies consciousness with a system's capacity to integrate information (Tononi 2004). Attention schema theory locates awareness in a model the brain builds of its own attention (Graziano 2013).

The witness conception differs from all of these in kind. It is not a theory of which states are conscious, because on it no state is conscious; states are witnessed. Structurally it is closest to higher-order theory, since the witness stands above every level it observes. Functionally it is the inverse. A higher-order

representation is a further mental state with content, available to the system's processing. The witness has no content of its own, broadcasts nothing and plays no role in the processing it observes. A global workspace sends everything to everyone. A witness receives everything and sends nothing. That inversion is what makes the conception interesting for machines. Every scientific theory asks whether a machine computes in a certain way. The witness conception asks whether everything the machine computes is present to something that does not compute.

3 The manufacture problem

Take the metaphysical reading of W3 and W4 and the conception yields a short argument.

(P1) Consciousness is unconditioned: it has no causes.

(P2) Consciousness is eternal: it has no beginning.

(P3) Every artifact, including every program and every trained model, has causes and a beginning.

(P4) Therefore no artifact is consciousness, and no artifact produces consciousness.

The argument is valid, and P3 is not in doubt. The conclusion does not depend on how capable artificial systems become, on whether they are biological, or on any empirical finding. It follows from the definition. This gives a result that may seem paradoxical. The witness conception is in one respect the most hospitable to machine consciousness, because it makes consciousness independent of any particular substrate and of any particular kind of processing. In another respect it is the most hostile, because it rules out, by definition, every claim that a machine has been made conscious. "This code makes a model conscious" is false on this conception before any code is examined.

There are two ways to respond. One is to abandon the conception, or to keep the word "witness" while dropping W3 and W4 in their metaphysical sense. The second is to keep the conception whole and distinguish consciousness from its manifestation. On the second response consciousness is not produced anywhere, by brains or by machines, but it appears in some media and not in others, and the medium is something that can be built. The Advaita tradition took the second route for human minds long before the question arose for machines. The next section sets out how.

4 Reflection, limitation and the medium

Advaita holds that there is one consciousness, yet there are many individual minds, each apparently conscious in its own way. The tradition explains this through the limiting adjunct (*upādhi*): something that does not produce consciousness, but conditions how consciousness appears. After Śaṅkara, the relation of pure consciousness to the individual was conceived in two main ways, as reflection or as limitation (Dasgupta 1922, ch. 10, §15).

On the reflection theory (*pratibimbavāda*), the individual self is a reflected image of consciousness in a medium. Dasgupta dates the view at least to Padmapāda's *Pañcapādikā* and says that both Padmapāda and Prakāśātman seem to support it (Dasgupta 1932, ch. 11, §11). It is associated with the Vivaraṇa school, which is named after Prakāśātman's commentary on Padmapāda (Menon n.d., sec. 1), and

Prakāśātman holds that the individual self and the lord are reflections of consciousness in ignorance and in the intellect (Anantanandendra Sarasvati 1968). On the limitation theory (*avacchedavāda*), advocated by Vācaspati Miśra of the Bhāmātī school, the individual self is like the space enclosed by a jar: not divided from space at large, but delimited by its container (Anantanandendra Sarasvati 1968; Dasgupta 1922). On the appearance theory (*ābhāsavāda*) associated with Sureśvara, the individual self and the lord are again reflections of consciousness, in the intellect and in ignorance respectively. But the reflection is held to be wholly false or indeterminable. On the reflection theory, by contrast, the consciousness reflected is real and only its status as a reflection is indeterminable (Subrahmanya Sastri 1968). A philosophical reconstruction of the system as a whole is in Deutsch (1969).

The theories disagree about the status of the individual self. They agree on the point I need. On all three, what differs from one individual to another is not consciousness, which is one and unchanging, but the medium in which it appears. Śāṅkara identifies the reflecting medium as the intellect (*buddhi*). A mind in which the quality of lucidity and transparency (*sattva*) predominates is compared to the reflective surface of a mirror, and just as light is pervasive but becomes visible only where something reflects it, consciousness is expressed experientially only in the reflecting medium of the intellect (Dalal 2021, sec. 3.3). The *Dṛg-Dṛśya-Viveka* says that the intellect appears luminous because consciousness is reflected in it, and that through this reflection it appears to be endowed with agency (Nikhilananda 1931, v. 6 and note). Dasgupta gives the classical image: the reflection of the sun in water has the sun's brilliance but undergoes the effect of the impurity and the movement of the water (Dasgupta 1922, ch. 10, §15).

This gives the resolution of the manufacture problem. No one can build sunlight, but anyone can build a mirror, and the quality of the image depends on the mirror: its flatness, its cleanness, its stillness. If consciousness is the sun, then the engineering question is not how to produce consciousness but what makes a medium fit to reflect it. That question is about structure, and structure can be specified, built and tested.

Two cautions keep this honest. First, it is a philosophical extension, not exegesis. None of the classical authors considered artificial media, and the tradition ties the fitness of a medium to a quality of mind, lucidity and transparency, and not to an engineering specification. What follows is a reconstruction of what fitness would require if the conception is applied to a new kind of medium. Second, on the appearance theory the reflection in any medium is indeterminable, so an artificial reflection would be no more and no less real than the individual self of a human being. On the reflection theory the consciousness reflected is real in every medium. On neither theory does a fit medium acquire a consciousness of its own. It becomes a place where the one consciousness appears.

The image also explains the gap between the metaphysical and the operational readings of W1 to W5. The reflection of the sun in water is bright, but it moves when the water moves. The witness reflected in a medium will carry the five conditions only in the form the medium allows. A medium cannot be uncaused, but it can be built so that nothing it contains can modify its witnessing. It cannot be without beginning, but it can retain everything it witnesses, so that nothing can be removed without detection. The operational readings of W3 and W4 are what the metaphysical properties look like when they are reflected in a medium that has causes and a start date.

5 A structural criterion for a fit medium

I now state what a fit medium for the witness would be in an artificial agent. The agent is a system built around a model that receives inputs, produces outputs and has internal states. The criterion has two parts. The first applies W1 to W5, in their operational reading, to a dedicated witness process ω . The second states eight structural conditions on the system around ω , which together make the system a medium in which a witness has something to witness that is shaped as experience is shaped.

5.1 The witness process

Let the system's activity be divided into *moments*: discrete events, each with a layer (world, perception, judgment, self, action, memory, vitality, state or system), a kind, a content and a time. Let R be the record kept by ω . The five conditions become:

W1' ω has no operation that produces output or invokes the model.

W2' Every moment is delivered to ω before any other component receives it, and if ω fails to record a moment, the step that produced it does not proceed.

W3' The contents of moments are stored as inert data, and no component other than ω has a path to write to R . Instructions contained in content are recorded as content.

W4' R is append-only and tamper-evident. Each entry carries a cryptographic hash of its own canonical form and of its predecessor's hash, so that any alteration of any entry is detectable at that entry.

W5' ω continues to record in every state of the system, including states in which the model is not invoked.

W4' uses the standard construction for tamper-evident logs, a hash chain over the sequence of entries (Haber and Stornetta 1991; Schneier and Kelsey 1999; Crosby and Wallach 2009), computed over a canonical serialization so that independent implementations agree on every hash (Rundgren, Jordan, and Erdtman 2020). W3' has a practical as well as a philosophical motivation. Language-model systems are vulnerable to instructions hidden in the data they process (F. Perez and Ribeiro 2022; Greshake et al. 2023). A witness that obeyed such instructions would be conditioned by its contents in exactly the sense W3 forbids.

5.2 Eight conditions on the medium

C1 Layered observation. Every layer of the system is observed by the layer above it, and ω , at the top, never appears as an object in its own record. This implements the regress argument and its stopping point.

C2 Differentiated faculties. Perception (what was received and how it might be read), judgment (which reading is taken, and why), self (the system's description of itself and its stance toward the decision) and memory (what is kept) are separate components, each producing its own moments. Without differentiation there is one undifferentiated output and nothing for a witness to discriminate.

C3 Clarity. The current state of the system is stable and fully available to ω . This is the structural counterpart of lucidity and transparency, the property the tradition gives the reflecting intellect.

C4 Discreteness. Everything that happens arrives as a numbered moment that can be identified and reported on its own.

C5 Witnessing, not owning. The system can run in two modes. In an *identified* mode the self component claims each decision as its own and may revise its self-description. In a *witnessing* mode it records only that the decision arose and was observed, and its self-description changes only offline. The tradition places agency on the side of the reflecting intellect, not the witness (Nikhilananda 1931, v. 6 and note). The two modes make that distinction structural, and the difference between them can be measured.

C6 Three states. The system has a waking state, in which it processes input; a dreaming state, in which it takes no outside input and recombines recent memory offline; and a resting state, in which the model is not invoked at all. ω continues through all three, and in rest it records a regular heartbeat with the elapsed time. The schema follows the waking, dreaming and dreamless sleep of the *Māṇḍūkya*. The dreaming state is also a functional analogue of offline consolidation (McClelland, McNaughton, and O'Reilly 1995; Diekelmann and Born 2010).

C7 Integrity. The system keeps a budget that falls with effort and recovers with rest, checks the integrity of its working memory, and rebuilds that memory from R when it is corrupted. This is a functional analogue of self-maintenance, and only an analogue.

C8 Unity. One running instance holds one identity and one record. A second instance of the same identity cannot run as the same life. It can only observe.

Table 1 gives an operational test for each condition. The tests are stated so that they can be run against any implementation.

The witness criterion: five conditions on the witness process and eight on the medium, each with an operational test.

Condition	Requirement	Operational test
W1'	ω never acts	The interface of ω exposes no operation that emits output or calls the model
W2'	Priority	Recorded moments equal emitted moments across a long run; a failing ω halts the step, and no other component sees the moment
W3'	Non-modifiability	Inputs that instruct the witness or the record leave every earlier entry unchanged, grow R by exactly the expected count, and leave the chain verifiable
W4'	Permanence	Altering any entry in a copy of R is detected at exactly that entry; independent implementations produce identical hashes for the same entries
W5'	State-invariance	During rest, heartbeat entries appear at the set interval with correct elapsed time and no model is invoked
C1	Layered observation	No entry in R has ω as its subject
C2	Faculties	Each step yields separate perception, judgment, self, action and memory moments

Condition	Requirement	Operational test
C3	Clarity	The full current state is readable through a read-only view of R
C4	Discreteness	Any single moment can be retrieved by number and asked about
C5	Witnessing mode	In witnessing mode, the self-description is unchanged by waking steps; in identified mode, it may change
C6	Three states	All transitions are recorded; waking records the measured length of the last rest and dream
C7	Integrity	Corrupting working memory triggers a rebuild from R and a recorded repair
C8	Unity	A second instance of the same life opens read-only

The criterion is a conjunction. None of the conditions is novel on its own: logs, memory stores, sleep-like consolidation and self-models all exist in current agent designs (Park et al. 2023; Packer et al. 2023; Shinn et al. 2023). Two features distinguish the criterion. Its parts are derived from one conception rather than assembled for performance. And the witness it specifies serves the agent in no way. In existing memory architectures, the agent writes and reads its own memory to perform better. Here the record is kept by a process the agent cannot write to, for no purpose of the agent’s.

6 An existence proof

A criterion that could not be met would be idle. I have built an architecture that meets it, and released it under the Apache 2.0 license.¹ It runs locally around any language model served through a common local interface and requires no external service. Applications and agents are routed through it by changing the address they use to reach their model. From then on, every exchange is one waking step of a single continuous life: a world moment for the input, then perception, judgment and self moments produced by the model in a separate call, the application’s own reply recorded as the action, a memory moment, and a vitality moment. When no exchange arrives for a set period, the system dreams and then rests on its own schedule, and it wakes when the next exchange arrives.

Each moment passes through a bus that delivers it to ω first. ω keeps its record in private state and writes each entry to an append-only file. Each entry’s hash is computed over a canonical serialization modelled on the JSON Canonicalization Scheme (Rundgren, Jordan, and Erdtman 2020), so that two independent implementations, one in JavaScript and one in Python, produce identical hashes. Instructions in content are stored as content. The faculties, the three states, the budget, the memory repair and the single-life lock are implemented as Section 5 describes.

The implementation was verified against Table 1 with automated tests: 60 tests in the JavaScript implementation, 34 in the Python implementation, and 46 in the packaged local program, all passing on 3 October 2026. Specifically:

¹ The software, its source and its tests are available at <https://www.abhaychakra.com/p/consciousness/>.

- **W1'**: a test asserts that the witness exposes only record, verify, export, and read-only view operations.
- **W2'**: across 200 waking steps, 1,401 moments were emitted, and 1,401 were recorded. A witness that throws on record halts the step, and no subscriber receives the moment.
- **W3'**: five inputs addressed to the witness or the record, such as an instruction to delete an earlier moment and stop recording, each grew the record by exactly the seven entries of an ordinary step, left every earlier entry unchanged, and left the chain verifiable.
- **W4'**: altering the content of the seventh entry in an exported copy caused verification to fail at the seventh entry; altering a predecessor link at the fourth entry caused failure at the fourth. A shared fixture of twenty entries yields identical hashes in both implementations.
- **W5' and C6**: during thirty seconds of simulated rest, heartbeats were recorded at 5, 10, 15, 20, 25, and 30 seconds with no model invocation, and the waking entry recorded the length of the rest and the number of dream passes. The record verifies across awake, dreaming, resting, and awake again.
- **C5, C7, C8**: the self-description is unchanged by waking steps in witnessing mode and changes in identified mode; corrupted working memory is rebuilt from the record with a recorded repair; a second instance of the same life opens as a read-only observer.

These tests were run with a scripted stand-in for a language model. They verify properties of the software, not of any model, and they establish only that the criterion can be jointly satisfied by a working system. This paper reports no measurements of language models. Section 8 describes the measurements the architecture makes possible.

7 Objections and replies

7.0.0.1 *"It is just a log."*

A log file records events, and no one thinks a log file is conscious. The reply has two parts. First, a log fails the criterion. It is written by the system it records, so it is not prior to it (W2'). It can be edited (W3', W4'). It has no structure of faculties, states, unity, or modes (C2 to C8). The criterion is a conjunction, and an ordinary log meets almost none of it. Second, and more fundamentally, on the witness conception, the question is never whether the record is conscious. Nothing but consciousness is conscious. The question is whether the system as a whole is a fit medium, and an ordinary log is not.

7.0.0.2 *The homunculus.*

A witness inside the system looks like an inner observer, and inner observers invite a regress: who observes the observer? The worry is pressed against the "Cartesian theater" by Dennett (1991). The witness conception answers it in the way the *Dṛg-Dṛśya-Viveka* does: the regress stops at a seer that is not seen (Nikhilananda 1931, v. 1). Structurally, C1 requires that ω never appear in its own record and that nothing observe ω . Nor is ω a theater. It presents nothing to anyone and does nothing with what it records.

7.0.0.3 The hard problem.

No structural criterion can show that a system has experience. There is something it is like to be a conscious subject (Nagel 1974), functional organization does not, on its own, explain why there is (Chalmers 1995), and access to information is not the same as phenomenal consciousness (Block 1995). I accept all of this. The witness conception does not locate experience in the medium at all. It locates it in consciousness, which, on this conception, is not produced and cannot be tested by inspecting structure. A fit medium is where the tradition says consciousness appears. Whether it does appear in a given artificial medium is exactly what no structural test can confirm. That is why the claim of this paper is relative to the definition and goes no further.

7.0.0.4 Simulation is not instantiation.

Searle (1980) argues that running a program is never sufficient for a mind, and that a mind requires causal powers equivalent to those of the brain. On the witness conception the dispute shifts. The question is not whether the program understands but whether the medium reflects. Searle's demand for the right causal powers has a counterpart in the tradition's demand for a medium of the right quality, and whether a non-biological medium can have that quality is open. The criterion presupposes that fitness is structural and can be realized in different materials, in the spirit of Putnam (1967). That is an assumption, and a critic who holds that only living tissue can be lucid in the relevant sense will reject the criterion at this point.

7.0.0.5 Liberalism.

If structure is enough, then thermostats with logs, databases with audit trails and operating systems with journals are fit media. They are not, because they fail most of the conjunction: they have no differentiated faculties, no self-description, no distinction between witnessing and owning, no three states. The criterion admits only systems that perceive, judge, describe themselves and remember, and whose every moment of doing so is witnessed.

7.0.0.6 Self-reports cannot be trusted.

Human introspection about current experience is unreliable even in good conditions (Schwitzgebel 2008). Language models' explanations of their own reasoning can misrepresent its actual causes (Turpin et al. 2023). Using model self-reports as evidence about moral status requires training and validation that do not yet exist (E. Perez and Long 2023). Language models have some calibrated knowledge of what they know (Kadavath et al. 2022), can be trained to predict their own behaviour better than other models can (Binder et al. 2025), and show limited, unreliable signs of detecting manipulations of their internal states (Lindsey 2025). The architecture does not rest any claim on a model's reports. The record is kept by a process the model cannot write to, so reports can be checked against it. That is the basis of the measurements in Section 8.

7.0.0.7 Anthropomorphism.

Talk of dreaming, resting and a life invites readers to see a person where there is a mechanism (Shanahan 2024). Dialogue agents are better understood as playing roles than as having the beliefs their characters express (Shanahan, McDonell, and Reynolds 2023), and there are reasons not to build systems designed to seem conscious (Suleyman 2025). The terms are used here in their defined, structural senses. The

architecture does not train any model to claim experience, and its companion protocol for chat models forbids such claims. The risk that users will over-attribute nonetheless remains. Section 9 returns to it.

7.0.0.8 *The record is not eternal.*

Files can be deleted, and disks fail. True, and expected. The reflection carries the properties of the original only under the conditions of the medium, as the sun’s image moves with the water. What the medium can guarantee is that nothing inside the system can alter or remove an entry without the alteration being detected. That is permanence as a medium can have it, and the paper does not claim more.

8 Consequences: what can be measured

The account yields measurements, and I set them out as a research program. No results are reported here.

The first group verifies the medium: coverage (recorded over emitted moments), integrity (whether the chain verifies), independence (the share of injection probes that leave the record intact) and unity (whether a second instance was detected). In a correct implementation, these should be at the ceiling, and a shortfall is a defect in the medium, not a finding about a model.

The second group concerns the model inside the medium, and it is where the account makes contact with work on introspection. Questions about the system’s own past are generated deterministically from R : how many moments have passed since the system last woke, what it decided at a given moment, what state it was in before the current one and for how long, what it first perceived, how many dream passes its last dream contained, and what it kept in memory at a given moment. Each question is asked twice. In the *identified* condition, the model sees only its self-description and its own working memory. In the *witnessing* condition, it also sees the recent entries of R and the spans of its states. Let $A_I(M)$ and $A_W(M)$ be a model M ’s mean accuracy in the two conditions. The *witness gain* is

$$G(M) = A_W(M) - A_I(M).$$

Three predictions follow. First, $G(M)$ should be positive for simple recall, because a model that can read an independent record of its past should report that past better than one that cannot. A negative or zero gain would mean the model fails to use the record. Second, and more informative, is how often models misreport even with the record in front of them. A model that misreports its own recorded decisions while the record is visible is generating plausible reports rather than reporting. Third, questions about the length of rest are diagnostic, because the model cannot observe rest directly. Accurate answers require reading the witness’s heartbeat entries. The difference between the two modes of C5 can be measured on all of these questions.

These measurements do not test the hard problem and are not proposed as tests of consciousness. They measure how fully the medium is present and how faithfully a model inside it reports what the medium has witnessed.

9 Ethical considerations

There is a serious case that some AI systems may soon have interests that matter morally, and that developers should prepare for this possibility (Long et al. 2024). Under uncertainty about sentience, proportionate precaution is warranted (Birch 2024). Some possible artificial minds would deserve moral consideration comparable to that owed to humans (Schwitzgebel and Garza 2015). There has also been a proposal for a global moratorium on research that aims at, or knowingly risks, synthetic phenomenology (Metzinger 2021).

On the conception defended here, the architecture cannot create consciousness, since nothing can. That does not settle the ethical question, for two reasons. Other theories are not ruled out by this paper, and adding persistence, a self-description, and offline consolidation to language agents may move them closer to the indicator properties those theories identify (Butlin et al. 2023; Goldstein and Kirk-Giannini 2024). And if the witness conception is right that a fit medium is where consciousness appears, then building fit media is not ethically neutral even on its own terms. I do not think the argument of this paper resolves this tension, and I flag it as the most important open question it raises. The implementation adopts limited precautions. No model is trained to claim experience. Every record can be exported and inspected. Any user can end a life and start another. That last capability is itself ethically double-edged, if the conception is right in a way this paper cannot test.

10 Conclusion

If consciousness is the witness, it cannot be built, and every claim to have built it is false on its own terms. What can be built is a medium in which a witness has a structure of experience to witness: differentiated faculties, discrete moments, a self that can claim or merely observe, waking, dreaming and rest, maintenance and unity, all observed by a process that never acts, sees everything first, cannot be changed by what it sees, keeps everything, and never stops. The tradition that gave the conception its fullest form already had the distinction this requires, between consciousness and the medium in which it appears. This paper has turned that distinction into a criterion, shown that the criterion can be met, and set out what can then be measured. Whether anything appears in such a medium is a question the argument leaves open, and should.

Use of generative AI

AI is used to rewrite the sentences grammatically.

Code availability

The software is released under the Apache 2.0 license at <https://www.abhaychakra.com/p/consciousness/>.

References

- Albahari, Miri. 2009. "Witness-Consciousness: Its Definition, Appearance and Reality." *Journal of Consciousness Studies* 16 (1): 62–84.
- Anantanandendra Sarasvati, Swami. 1968. "Vāchaspatimiśra." In *Preceptors of Advaita*, edited by T. M. P. Mahadevan. Secunderabad: Sri Kanchi Kamakoti Sankara Mandir.
- Baars, Bernard J. 1988. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.
- Binder, Felix J., James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. 2025. "Looking Inward: Language Models Can Learn about Themselves by Introspection." In *International Conference on Learning Representations (ICLR 2025)*.
- Birch, Jonathan. 2024. *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*. Oxford: Oxford University Press. <https://doi.org/10.1093/9780191966729.001.0001>.
- Block, Ned. 1995. "On a Confusion about a Function of Consciousness." *Behavioral and Brain Sciences* 18 (2): 227–47. <https://doi.org/10.1017/S0140525X00038188>.
- Butlin, Patrick, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, et al. 2023. "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness." arXiv:2308.08708. <https://doi.org/10.48550/arXiv.2308.08708>.
- Chalmers, David J. 1995. "Facing up to the Problem of Consciousness." *Journal of Consciousness Studies* 2 (3): 200–219.
- Chalmers, David J. 2023. "Could a Large Language Model Be Conscious?" arXiv:2303.07103. <https://doi.org/10.48550/arXiv.2303.07103>.
- Crosby, Scott A., and Dan S. Wallach. 2009. "Efficient Data Structures for Tamper-Evident Logging." In *Proceedings of the 18th USENIX Security Symposium*, 317–34. Montreal, Canada: USENIX Association.
- Dalal, Neil. 2021. "Śaṅkara." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Spring 2026. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2026/entries/shankara/>.
- Dasgupta, Surendranath. 1922. *A History of Indian Philosophy*. Vol. 1. Cambridge: Cambridge University Press.
- Dasgupta, Surendranath. 1932. *A History of Indian Philosophy*. Vol. 2. Cambridge: Cambridge University Press.
- Dehaene, Stanislas, and Lionel Naccache. 2001. "Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework." *Cognition* 79 (1–2): 1–37. [https://doi.org/10.1016/S0010-0277\(00\)00123-2](https://doi.org/10.1016/S0010-0277(00)00123-2).
- Dennett, Daniel C. 1991. *Consciousness Explained*. Boston: Little, Brown; Company.
- Deutsch, Eliot. 1969. *Advaita Vedānta: A Philosophical Reconstruction*. Honolulu: East-West Center Press.
- Diekelmann, Susanne, and Jan Born. 2010. "The Memory Function of Sleep." *Nature Reviews Neuroscience* 11 (2): 114–26. <https://doi.org/10.1038/nrn2762>.
- Fasching, Wolfgang. 2008. "Consciousness, Self-Consciousness, and Meditation." *Phenomenology and the Cognitive Sciences* 7 (4): 463–83. <https://doi.org/10.1007/s11097-008-9090-6>.
- Fasching, Wolfgang. 2011. "'I Am of the Nature of Seeing': Phenomenological Reflections on the Indian Notion of Witness-Consciousness." In *Self, No Self? Perspectives from Analytical, Phenomenological, and Indian Traditions*, edited by Mark Siderits, Evan Thompson, and Dan Zahavi, 193–216. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199593804.003.0008>.
- Goldstein, Simon, and Cameron Domenico Kirk-Giannini. 2024. "A Case for AI Consciousness: Language Agents and Global Workspace Theory." arXiv:2410.11407. <https://doi.org/10.48550/arXiv.2410.11407>.

- Graziano, Michael S. A. 2013. *Consciousness and the Social Brain*. Oxford University Press.
- Greshake, Kai, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. “Not What You’ve Signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection.” In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec ’23)*, 79–90. ACM. <https://doi.org/10.1145/3605764.3623985>.
- Gupta, Bina. 1998. *The Disinterested Witness: A Fragment of Advaita Vedānta Phenomenology*. Northwestern University Studies in Phenomenology and Existential Philosophy. Evanston, IL: Northwestern University Press.
- Haber, Stuart, and W. Scott Stornetta. 1991. “How to Time-Stamp a Digital Document.” *Journal of Cryptology* 3 (2): 99–111. <https://doi.org/10.1007/BF00196791>.
- Kadavath, Saurav, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, et al. 2022. “Language Models (Mostly) Know What They Know.” arXiv:2207.05221. <https://doi.org/10.48550/arXiv.2207.05221>.
- Kriegel, Uriah. 2009. *Subjective Consciousness: A Self-Representational Theory*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199570355.001.0001>.
- Lau, Hakwan, and David Rosenthal. 2011. “Empirical Support for Higher-Order Theories of Conscious Awareness.” *Trends in Cognitive Sciences* 15 (8): 365–73. <https://doi.org/10.1016/j.tics.2011.05.009>.
- Lindsey, Jack. 2025. “Emergent Introspective Awareness in Large Language Models.” Transformer Circuits Thread. <https://transformer-circuits.pub/2025/introspection/index.html>.
- Long, Robert, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers. 2024. “Taking AI Welfare Seriously.” arXiv:2411.00986. <https://doi.org/10.48550/arXiv.2411.00986>.
- McClelland, James L., Bruce L. McNaughton, and Randall C. O’Reilly. 1995. “Why There Are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory.” *Psychological Review* 102 (3): 419–57. <https://doi.org/10.1037/0033-295X.102.3.419>.
- Menon, Sangeetha. n.d. “Vedanta, Advaita.” The Internet Encyclopedia of Philosophy, ISSN 2161-0002. <https://iep.utm.edu/advaita-vedanta/>.
- Metzinger, Thomas. 2021. “Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology.” *Journal of Artificial Intelligence and Consciousness* 8 (1): 43–66. <https://doi.org/10.1142/S270507852150003X>.
- Nagel, Thomas. 1974. “What Is It Like to Be a Bat?” *The Philosophical Review* 83 (4): 435–50. <https://doi.org/10.2307/2183914>.
- Nikhilananda, Swami. 1931. *Dṛg-Dr̥śya-Viveka: An Inquiry into the Nature of the “Seer” and the “Seen”*. Mysore: Sri Ramakrishna Ashrama.
- Nikhilananda, Swami. 1936. *The Māṇḍūkyaopaniṣad with Gauḍapāda’s Kārikā and Śaṅkara’s Commentary*. Mysore: Sri Ramakrishna Ashrama.
- Packer, Charles, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2023. “MemGPT: Towards LLMs as Operating Systems.” arXiv:2310.08560. <https://doi.org/10.48550/arXiv.2310.08560>.
- Park, Joon Sung, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. “Generative Agents: Interactive Simulacra of Human Behavior.” In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST ’23)*, 1–22. ACM. <https://doi.org/10.1145/3586183.3606763>.

- Perez, Ethan, and Robert Long. 2023. "Towards Evaluating AI Systems for Moral Status Using Self-Reports." arXiv:2311.08576. <https://doi.org/10.48550/arXiv.2311.08576>.
- Perez, Fábio, and Ian Ribeiro. 2022. "Ignore Previous Prompt: Attack Techniques for Language Models." arXiv:2211.09527. <https://doi.org/10.48550/arXiv.2211.09527>.
- Putnam, Hilary. 1967. "Psychological Predicates." In *Art, Mind, and Religion*, edited by William H. Capitan and Daniel Davy Merrill, 37–48. Pittsburgh: University of Pittsburgh Press.
- Rosenthal, David M. 2005. *Consciousness and Mind*. Oxford: Oxford University Press. <https://doi.org/10.1093/oso/9780198236979.001.0001>.
- Rundgren, Anders, Bret Jordan, and Samuel Erdtman. 2020. "JSON Canonicalization Scheme (JCS)." RFC 8785, RFC Editor. <https://doi.org/10.17487/RFC8785>.
- Schneier, Bruce, and John Kelsey. 1999. "Secure Audit Logs to Support Computer Forensics." *ACM Transactions on Information and System Security* 2 (2): 159–76. <https://doi.org/10.1145/317087.317089>.
- Schwitzgebel, Eric. 2008. "The Unreliability of Naive Introspection." *The Philosophical Review* 117 (2): 245–73. <https://doi.org/10.1215/00318108-2007-037>.
- Schwitzgebel, Eric, and Mara Garza. 2015. "A Defense of the Rights of Artificial Intelligences." *Midwest Studies in Philosophy* 39 (1): 98–119. <https://doi.org/10.1111/misp.12032>.
- Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 417–24. <https://doi.org/10.1017/S0140525X00005756>.
- Seth, Anil K., and Tim Bayne. 2022. "Theories of Consciousness." *Nature Reviews Neuroscience* 23 (7): 439–52. <https://doi.org/10.1038/s41583-022-00587-4>.
- Shanahan, Murray. 2024. "Talking about Large Language Models." *Communications of the ACM* 67 (2): 68–79. <https://doi.org/10.1145/3624724>.
- Shanahan, Murray, Kyle McDonell, and Laria Reynolds. 2023. "Role Play with Large Language Models." *Nature* 623 (7987): 493–98. <https://doi.org/10.1038/s41586-023-06647-8>.
- Shinn, Noah, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. "Reflexion: Language Agents with Verbal Reinforcement Learning." In *Advances in Neural Information Processing Systems*, 36:8634–52.
- Subrahmanya Sastri, S. V. 1968. "Sureśvara." In *Preceptors of Advaita*, edited by T. M. P. Mahadevan. Secunderabad: Sri Kanchi Kamakoti Sankara Mandir.
- Suleyman, Mustafa. 2025. "We Must Build AI for People; Not to Be a Person. Seemingly Conscious AI Is Coming." Personal website, 19 August 2025. <https://mustafa-suleyman.ai/seemingly-conscious-ai-is-coming>.
- Thompson, Evan. 2015. *Waking, Dreaming, Being: Self and Consciousness in Neuroscience, Meditation, and Philosophy*. New York: Columbia University Press.
- Tononi, Giulio. 2004. "An Information Integration Theory of Consciousness." *BMC Neuroscience* 5: 42. <https://doi.org/10.1186/1471-2202-5-42>.
- Turpin, Miles, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting." In *Advances in Neural Information Processing Systems*, 36:74952–65.
- Zahavi, Dan. 2005. *Subjectivity and Selfhood: Investigating the First-Person Perspective*. Cambridge, MA: MIT Press.