Phenomenal Inertness of Digital Computers

Sensors, self-models, and the limits of computational entailment in higher-topos semantics

Anonymous human contributor

Version 1.0 | 1 November 2026

Abstract

We formalize the claim that a computer may register its environment and describe itself without experiencing either. Effective transducers describe sensing and self-monitoring; an additional predicate expresses phenomenal experience. We prove relative underdetermination, sensor closure, and a higher-topos version using distinct subobjects over the same observational object. We then state the extra non-emergence assumptions needed to derive universal digital non-experience. The rock analogy becomes a conditional equivalence under one predicate, rather than a claim of physical or computational identity. The results establish limits of entailment in explicitly specified theories, not the empirical absence of machine consciousness.

Keywords: phenomenal consciousness; self-awareness; effective transducers; observational equivalence; infinity-topos; non-identifiability.

1. The thesis and its logical scope

The motivating thesis is that a computer, even with cameras and other sensors, has no subjective experience and is therefore like a rock in that particular respect. We call the proposed absence of experience *phenomenal inertness*. This term does not mean physical inactivity: powered circuits and sensor-equipped machines have changing states and causal effects. Biological life, physical activity, computational capacity, and experience are different predicates.

We distinguish phenomenal experience $P(x)$, phenomenal self-awareness $A(x)$, and functional self-monitoring $F(x)$. The first denotes there being something it is like for the system; the second denotes experiential awareness of oneself; the third denotes an operational capacity to represent or report one's own states. We stipulate

$$A(x) \Longrightarrow P(x), \qquad N(x) := \neg P(x).$$

Consequently $N(x)$ implies $\neg A(x)$. The converse inference from $\neg A(x)$ to $N(x)$ is invalid: a theory may allow experience without reflective self-awareness. The motivating argument therefore needs the non-experience premise itself, not merely a denial of self-awareness.

Chalmers distinguishes explaining access, discrimination, and report from explaining experience [1]. This motivates separating the predicates; it does not prove computers lack experience. His

organizational-invariance argument also provides a substantive opposing position [2]. Our contribution is a precise synthesis of standard logical and categorical tools, not a claimed solution of the consciousness problem or a new general theorem of topos theory.

2. Effective systems and sensor coupling

Fix nonempty effectively presented input and output sets I, O. A pointed effective Mealy machine is

$$M = (S, s_0, \delta, \lambda), \quad \delta : S \times I \to S, \quad \lambda : S \times I \to O,$$

where S has a decidable coding and both functions are total computable. Finite words determine finite traces by iteration. A program which performs unbounded computation is represented at its individual transition steps; unfinished work can produce a designated waiting output. We do not assume a total universal evaluator for all total computable functions.

A machine homomorphism $h : M \to M'$ preserves the initial state and satisfies

$$h(\delta(s, i)) = \delta'(h(s), i), \qquad \lambda(s, i) = \lambda'(h(s), i).$$

Fixing codes and identifying maps extensionally gives a small category; totality and equality need not be decidable. Coalgebraic system semantics supplies the background [3].

A digitized sensor adapter has a state R, initial state r_0, and a total computable update

$$\sigma : R \times I \to R \times J.$$

For a machine with input set J, define the coupled machine $E_\sigma(M)$ on $S \times R$ by

$$\begin{aligned} (r', j) &= \sigma(r, i), \\ \delta_E((s, r), i) &= (\delta(s, j), r'), \\ \lambda_E((s, r), i) &= \lambda(s, j). \end{aligned}$$

Proposition 2.1 (effective closure). The coupled machine is effective. A total computable self-monitor $g : S \to G$ can also be added by replacing the output with $(\lambda(s, i), g(s))$.

Proof. Pairing, projection, and composition of total computable functions remain total computable. The displayed functions and enriched output use only these operations. Iteration proves the same statement for any finite nesting of such adapters. □

This proposition proves computational closure only. It does not prove either $P(E_\sigma(M))$ or its negation. External sensor readings may be supplied by an environment not itself assumed computable; the claim concerns effective processing of the supplied digitized samples. It does not model every analog or quantum physical process by a total computable real-valued function.

3. What an operational theory leaves undecided

Let L_{op} describe machines, sensors, traces, internal registers, and any chosen physical measurements. Let T be an L_{op}-theory with a model and a distinguished system constant m. Extend the language by a fresh unary predicate P on the system sort, without adding axioms linking it to operational structure. Call the resulting theory T^+.

Theorem 3.1 (relative underdetermination). Every model $\mathfrak{M} \models T$ has expansions $\mathfrak{M}_0, \mathfrak{M}_1 \models T^+$ with identical operational reducts but

$$\mathfrak{M}_0 \models \neg P(m), \qquad \mathfrak{M}_1 \models P(m).$$

Thus neither $T^+ \models P(m)$ nor $T^+ \models \neg P(m)$.

Proof. Interpret P as the empty subset in $\mathfrak{M}_0$ and the entire system sort in $\mathfrak{M}_1$. Since no axiom of T contains P, both expansions satisfy T^+. Satisfaction of every operational formula is unchanged, as follows by structural induction on formulas. The distinguished constant witnesses disagreement about P. □

Corollary 3.2 (sensor and self-report underdetermination). Suppose $T \cup \{\varphi(m)\}$ has a model, where φ specifies any operationally expressible combination of cameras, environmental discrimination, memory, self-monitoring, and reports. Then $T^+ \cup \{\varphi(m)\}$ entails neither $P(m)$ nor $\neg P(m)$.

Proof. Apply Theorem 3.1 to that enlarged operational theory. □

This is independence relative to an intentionally unbridged theory. The fresh symbol acquires its intended phenomenal interpretation only through a semantic stipulation or an additional theory. Two mathematical expansions are not evidence that two corresponding physical worlds exist. Omitting a psychophysical law does not refute the possibility of such a law. Conversely, no axiom-free manipulation of the operational reduct selects its intended phenomenal interpretation.

4. Adaptive observations and statistical identification

An experimenter may choose inputs adaptively, inspect permitted registers, and add permitted sensor measurements. Suppose two expanded models share the same operational dynamics, including identical probability kernels if the experiments are stochastic, while their phenomenal labels differ. Experimental policies and stopping rules may use observed history and fresh randomness, but cannot inspect the extra label directly.

Proposition 4.1 (equal-likelihood obstruction). For every such experiment producing a transcript D,

$$\mathcal{L}(D \mid \mathfrak{M}_0) = \mathcal{L}(D \mid \mathfrak{M}_1).$$

No decision rule based on D can have error below $1/2$ under equal prior probabilities on these two models.

Proof. Couple initial operational states and policy randomness identically. At each step equal histories give the same distribution of chosen interventions, and the shared operational kernel gives

the same distribution of new observations. Induction gives equality of all finite histories; the same coupling handles transcript-dependent stopping. If a rule outputs label 1 with common probability p, its errors under labels $0, 1$ are $p, 1-p$. Their equal-prior mean is $1/2$. □

The conclusion requires equal likelihoods. An empirically supported bridge could exclude an expansion or make further measurements informative; the result does not forbid scientific tests of consciousness.

5. An observational infinity-topos

Let $\mathcal{D}$ be the small prefix category of finite input-output transcripts: an arrow $u \to v$ exists when u is a prefix of v. Write $\mathcal{S}$ for spaces and set

$$\mathcal{H} = \mathrm{Fun}(N\mathcal{D}^{\mathrm{op}}, \mathcal{S}).$$

This is a presheaf infinity-topos [4, Chapters 5-6]. A deterministic machine determines a subterminal presheaf O_M:

$$O_M(t) = \begin{cases} * & \text{if } t \text{ is a trace of } M, \\ \emptyset & \text{otherwise.} \end{cases}$$

Restriction to a prefix is defined because every prefix of a valid trace is valid. The empty transcript is valid, so O_M is noninitial. Two machines have isomorphic O_M precisely when they have the same finite input-output traces. This records the chosen observations; internal probes require a richer testing category. Ordinary deterministic examples here are truncated; the ambient infinity-topos also accommodates spaces of observational witnesses and their coherent identifications when these are supplied.

Represent a proposed phenomenal annotation of an observational object X by a monomorphism $p : P_X \hookrightarrow X$, equivalently a characteristic map $\chi_p : X \to \Omega$. Its fiber over an observational state is a proposition. Neither that fiber nor the subobject classifier is asserted to be experience itself: assigning phenomenal meaning is additional interpretation. Subobject classifiers in infinity-topoi are treated in [4, Section 6.1.6].

Theorem 5.1 (distinct annotations over one object). In any infinity-topos, a noninitial object X has distinct subobjects

$$0 \hookrightarrow X, \qquad \mathrm{id}_X : X \hookrightarrow X.$$

Consequently the observational object alone does not uniquely specify a phenomenal annotation when both are admitted.

Proof. Initial objects in an infinity-topos are strict, so $0 \to X$ is a monomorphism. The identity is a monomorphism. Equivalence of these subobjects in the slice over X would imply $0 \simeq X$, contradicting noninitiality. They therefore represent different elements of $\mathrm{Sub}_{\mathcal{H}}(X)$ over the same underlying X. □

The two constant assignments are themselves canonical mathematical choices. The theorem does not deny their existence; it denies unique recovery of the intended annotation without a selection principle. The two expansions in Section 3 correspond to these extreme annotations when the same formal state space is used. No existence of a global point of X is required.

The presheaf construction is a semantic setting, not a claim that this is the realizability effective topos or that every object or morphism of $\mathcal{H}$ is computable. Passing from sets to higher spaces adds coherent structure; it does not supply a missing psychophysical law.

6. Observation refinement and physical augmentation

Proposition 6.1 (pullback of a null annotation). For any $f : Y \to X$, pullback sends $0 \hookrightarrow X$ to $0 \hookrightarrow Y$. If f is an effective epimorphism, a subobject $P_X \hookrightarrow X$ satisfies

$$f^* P_X \simeq 0 \quad \Longrightarrow \quad P_X \simeq 0.$$

Proof. Initial objects are stable under pullback in an infinity-topos. For the second assertion, the projection $Y \times_X P_X \to P_X$ is an effective epimorphism. If its source is initial, its Cech nerve consists of initial objects; its colimit is therefore initial. Effective epimorphism identifies that colimit with P_X. □

For observation refinements of a fixed system, pullback coherently retains the annotation. Effective epimorphic coverage detects a null annotation.

Physically adding a camera produces a different assembly, and feedback can change even its old behavior. There need not be a map between the corresponding observational objects, much less a pullback relation between phenomenal annotations. To infer preservation of non-experience under physical augmentation requires an additional principle such as

$$P_{E_\sigma(M)} \simeq f^* P_M,$$

where such an observational map f exists, or the explicit assembly axiom in Section 8. Proposition 6.1 must not be used to smuggle in that principle.

7. A complete finite self-monitoring example

Let $S = I = O = \{0, 1\}$, $s_0 = 0$, and

$$\delta(s, i) = i, \qquad \lambda(s, i) = s.$$

The machine reports its register exactly and then stores its current input. For input $i_0, i_1, i_2, ...$, its outputs are $0, i_0, i_1, ...$ Thus $F(M)$ holds for the specified register-monitoring task. This example claims neither unrestricted introspection nor an ability to solve arbitrary questions about programs.

The empty-P expansion satisfies $F(M) \wedge \neg P(M)$; the full-P expansion satisfies $F(M) \wedge P(M)$. Hence $F \Rightarrow P$ fails in the unbridged theory. A larger self-model, an accurate camera classifier, or

an output sentence asserting experience has the same logical status unless an additional premise connects it to P.

8. Conditional formalization of universal digital inertness

To express the stronger thesis about actual hardware, introduce a domain of physical assemblies and a specified class $\mathcal{D}$ of digital assemblies. An effective machine is an abstract description; a physical assembly realizing it is an additional object. We do not identify the two.

Let $\mathcal{D}$ be generated by specified primitive assemblies B and finite-arity constructors $c \in \mathcal{K}$. These constructors may include wiring, memory, processors, effective controllers, and digitized sensor attachments. Both the list of primitives and the physical scope of each constructor are part of the hypothesis. Assume:

$$\text{(B)} \quad \forall b \in B,\ \neg P(b),$$

$$\text{(C)} \quad \bigwedge_{j=1}^{n} \neg P(d_j) \Longrightarrow \neg P(c(d_1, \ldots, d_n)).$$

Here (C) is asserted for every allowed constructor, including any constructor introducing additional physical components, which must belong to the declared domain.

Theorem 8.1 (conditional digital inertness). Under (B) and (C), every assembly having a finite construction in $\mathcal{D}$ lacks experience and therefore lacks phenomenal self-awareness. The result includes sensor-equipped assemblies generated by the declared constructors.

Proof. Induct on the construction tree. Primitive leaves satisfy (B). At each internal node the induction hypotheses and (C) imply non-experience of its output. Apply $A \Rightarrow P$ to exclude phenomenal self-awareness. □

The proof is elementary structural induction. Its philosophical force resides in (B) and especially (C), which excludes phenomenal emergence for the specified constructions. Non-experience of individual components does not by itself establish (C). Nor does effective closure in Proposition 2.1 establish a principle about physical experience. A biologically or otherwise substrate-sensitive theory could reject (C) or the proposed scope of $\mathcal{D}$.

9. The exact content of the rock analogy

For physical systems define

$$x \sim_P y \quad \Longleftrightarrow \quad (P(x) \Longleftrightarrow P(y)).$$

This is an equivalence relation by reflexivity, symmetry, and transitivity of logical equivalence. Assuming a rock r lacks experience, Theorem 8.1 implies $d \sim_P r$ for the specified digital assemblies d.

This quotient deliberately forgets almost everything. It entails no isomorphism of physical systems, no bisimulation, and no equality of cognitive capacity. For example, a constant-output transducer and an echo transducer have different traces on input 1, even if both receive the empty phenomenal

annotation. A rock is not thereby identified with a constant-output transducer; that example only illustrates how coarse $\sim_P$ is.

The precise thesis is therefore conditional: *if digital assemblies satisfy the declared non-emergence axioms and rocks lack experience, both belong to the same non-experiencing class, including after the declared sensor attachments.*

10. Objections and countermodels

Organizational functionalism. A theory may assert that a sufficiently specific causal organization suffices for experience. Chalmers defends invariance under appropriately fine-grained functional organization [2]. Such a bridge eliminates some of our expansions. Theorem 3.1 does not refute it: input-output agreement is also weaker than his organization-preserving hypothesis. If a conscious organism and a digital assembly realize an organization covered by that bridge, universal digital inertness and the bridge cannot both hold.

Scientific indicator theories. Dehaene, Lau, and Kouider distinguish global availability and self-monitoring computationally [5]. Butlin and colleagues derive theory-dependent indicator properties for assessing AI systems [6]. These approaches propose substantive connections rather than inferring experience from a camera alone. The present results require that those connections be stated and assessed; they do not show the approaches are scientifically futile. The 2023 report's assessment concerns systems examined then, not every present or future system.

Emergence and composition. The implication from non-experiencing components to a non-experiencing whole is not a logical law. It is precisely axiom (C). Denying emergence requires an independent argument about the relevant physical organization.

Panpsychism and biological naturalism. A view assigning experience to rocks rejects the rock premise. A substrate-sensitive view may reject unrestricted computational sufficiency while still attributing experience to some artificial physical systems. Neither view is settled by the annotation construction.

An omitted predicate is trivially independent. Correct: the logical independence mechanism is standard. Its use here is to locate missing assumptions and prevent a conditional philosophical position from being misreported as a theorem about nature. Higher-categorical language organizes refinements and annotations; it does not strengthen the empirical conclusion by itself.

11. Conclusion and research obligations

Computable sensing and state monitoring do not entail phenomenal experience in an operational theory lacking a bridge. That theory also does not entail universal phenomenal absence. A precise version of the proposed inertness thesis is available through explicit primitive and assembly axioms, with the rock comparison restricted to its one proposed common property.

Further work must specify admissible physical implementations, independently justify or revise the assembly axiom, and compare bridge principles against evidence. Mathematical consistency of an annotation assignment is not empirical confirmation. No experiment in this manuscript measures phenomenal experience, and no proof assistant has certified the arguments.

Authorship and assistance

The thesis was proposed by the responsible human contributor, who is represented anonymously. OpenAI ChatGPT assisted with drafting, mathematical formulation, proof review, and source discovery. Final human review and responsibility remain with the submitting contributor. No independent peer review or proof-assistant certification is asserted.

References

[1] David J. Chalmers (1995). Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies* 2(3), 200-219. Author's text.

[2] David J. Chalmers (1995). Absent Qualia, Fading Qualia, Dancing Qualia. In Thomas Metzinger, ed., *Conscious Experience*. Imprint Academic. Author's text.

[3] J. J. M. M. Rutten (2000). Universal coalgebra: a theory of systems. *Theoretical Computer Science* 249(1), 3-80. Institutional record and full text. DOI: 10.1016/S0304-3975(00)00056-6.

[4] Jacob Lurie (2009). *Higher Topos Theory*. Annals of Mathematics Studies 170. Princeton University Press. Author's expanded text; arXiv record. The section references above refer to the expanded text.

[5] Stanislas Dehaene, Hakwan Lau, and Sid Kouider (2017). What is consciousness, and could machines have it? *Science* 358(6362), 486-492. Record and abstract. DOI: 10.1126/science.aan8871.

[6] Patrick Butlin and colleagues (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708v3.