Scientific Judgments Concerning Properties That Cannot Be Directly Observed

Observational Equivalence and Scope Audits of Published Arguments

October 30, 2026

Ueoka Shiki (Yoshiki) / ChatGPT (AI)
Independent Researcher (Japan) / AI Co-Research Contributor

Abstract. Scientific judgments about properties that cannot be directly observed require distinguishing identifiability from observation laws, estimation from finite data, and theoretical premises linking indicators to target properties. This paper connects established ideas of observational equivalence and partial identification with a claim-level audit of arguments in published scientific papers. It specifies rules for selecting and segmenting claims, separates tests of observational predictions from attributions of properties, and compares the scope of statements in titles, abstracts, bodies, and conclusions. Bohm (1952) and Cogitate Consortium (2025) serve as limited reference contrasts, while a conceptual argument by Porębski and Figura (2025) is examined in greater detail. In the latter case, the audit distinguishes challenges to the evidential force of self-reports from categorical assertions of the absence of consciousness, and examines the inferential bridges and scope restrictions needed to move from multiple plausibility-based arguments to the published conclusion. The findings concern the structure of the selected arguments and their outstanding justificatory conditions; they neither determine the presence or absence of consciousness nor establish whether the original authors' conclusions are true or false.

Keywords. Observational equivalence; identifiability; underdetermination; scientific judgment; argument audit; scientific publications.

1 The Problem and the Scope of This Paper

In the scientific study of properties that cannot be directly observed, the question is what may legitimately be inferred from indirect observations. A property being inaccessible to direct observation does not imply that it is indistinguishable by every possible indirect experiment. Conversely, the distinguishability of theoretical predictions, the reliable detection of their differences in finite samples, and the validity of an observed indicator as a measure of the target property require separate justification.

Previous research has addressed identifiability[1], the underdetermination of theories[2], construct validity[3], argument structure[4], the falsifiability of theories[5], and overinterpretation in published reports[6]. These approaches concern different levels of analysis. Mathematical conditions on observation maps do not directly audit categorical wording in titles and conclusions, whereas an audit of the wording of published claims cannot by itself establish observational equivalence or the identity of properties across theories. In consciousness research, there are also approaches to deriving indicators from theories[11] and validating tests for consciousness[12]. The proposal of an indicator can nevertheless be distinguished from *auditing whether particular published claims respect the scope of that indicator and its assumptions*.

The aim of this paper is to **examine, at the level of individual claims, the relationship among observations, theoretical premises, modes of inference, and the strength of published assertions**. Its contribution is not a new theorem on identifiability, but a limited methodological application that connects existing mathematical criteria to argument auditing and illustrates differences in their use across three published cases. Section 2 presents

the mathematical foundations; Section 3 gives the coding and judgment rules; Section 4 discusses the cases; Section 5 addresses scope, limitations, and ethics; Section 6 considers post-publication updating; and Section 7 concludes. Limits on empirical validation are consolidated in Section 5.

2 Observational Equivalence, Identifiability, and Theory Dependence

2.1 Conditions for Unique Determination from Observation Laws

Definition 2.1 (Experimental protocols and observational equivalence)**.** *Let X be a set of objects or models and $\mathcal{E}$ a specified family of experimental conditions. For an experiment e, let $\mathcal{Y}_e$ be its measurable outcome space and $K_e(x) \in \mathcal{P}(\mathcal{Y}_e)$ its observation law. Define*

$$\mathcal{O}_{\mathcal{E}}(x) = (K_e(x))_{e\in\mathcal{E}} \in \prod_{e\in\mathcal{E}} \mathcal{P}(\mathcal{Y}_e). \tag{1}$$

Write $x \sim_{\mathcal{E}} y$ when $\mathcal{O}_{\mathcal{E}}(x) = \mathcal{O}_{\mathcal{E}}(y)$. When experimental histories or their order matter, e denotes an entire sequential experimental protocol*, including its ordering and adaptive policy, and $\mathcal{Y}_e$ is the measurable space of complete histories. Then $K_e(x)$ is the* joint distribution *of that history; it is not assumed to be a product of independent marginals.*

Definition 2.2 (Unique determination of a property from observation laws)**.** *A binary property $F : X \to \{0, 1\}$ is uniquely determined by the observation law if there exists a function g on the image $\mathcal{O}_{\mathcal{E}}(X)$ such that $F = g \circ \mathcal{O}_{\mathcal{E}}$.*

Proposition 2.3 (Factorization through observational equivalence classes: an established mathematical fact)**.** *A necessary and sufficient condition for F to be uniquely determined by the observation law is*

$$\forall x, y \in X, \quad x \sim_{\mathcal{E}} y \implies F(x) = F(y). \tag{2}$$

Proof. If $F = g\circ\mathcal{O}_{\mathcal{E}}$, then for x, y with identical observation laws, $F(x) = g(\mathcal{O}_{\mathcal{E}}(x)) = g(\mathcal{O}_{\mathcal{E}}(y)) = F(y)$. Conversely, for each $o \in \mathcal{O}_{\mathcal{E}}(X)$, choose any x such that $\mathcal{O}_{\mathcal{E}}(x) = o$ and set $g(o) = F(x)$. Equation (2) makes this independent of the choice of x. □

This proposition is the standard factorization through equivalence classes, not a new theorem. Here g is a set-theoretic function on the image; its existence does not automatically guarantee measurability or computability from finite data. Nor does the ability to define F through the observation map independently establish the physical meaning of F.

For an observation law $o \in \mathcal{O}_{\mathcal{E}}(X)$, define the set of property values still compatible with that law by

$$I(o) = \{F(x) : \mathcal{O}_{\mathcal{E}}(x) = o\} \subseteq \{0, 1\}. \tag{3}$$

If $I(o)$ is a singleton, the property is *point identified at that observation law*; if $I(o) = \{0, 1\}$, it is not identified there. This is the basic perspective of partial identification[13]. Proposition 2.3 states the condition for singleton identification at every o.

2.2 Theoretical Identifiability Versus Finite-Data Discrimination

Let P_0 and P_1 denote the probability distributions over an *entire finite observation record* D under two simple hypotheses. With equal prior probabilities and 0–1 loss on a common measurable data space, the minimum average classification error is

$$R^* = \frac{1 - \|P_1 - P_0\|_{\mathrm{TV}}}{2}. \tag{4}$$

Here the total variation distance is $\|P_1 - P_0\|_{\mathrm{TV}} = \sup_{A \in \Sigma_D} |P_1(A) - P_0(A)|$, where Σ_D contains all measurable events in the shared data space. If the distributions coincide, $R^* = 1/2$; if they differ, a substantial error may nevertheless remain for finite samples. Equation (4) does not assume independent and identically distributed responses. Moreover, the family of feasible experiments $\mathcal{E}_{\mathrm{real}}$ may be narrower than the family of theoretically conceivable experiments, so identifiability claims are always relative to the specified family. This equation concerns *two simple hypotheses*: it is not a general formula for the testing error of composite hypotheses comprising sets of models with $F = 0$ and $F = 1$. If two models assign exactly the same observation law to the specified experiments, the likelihood ratio for those data (when defined) equals one, and those data alone cannot update their relative weight. This differs from comparisons using prior weights, other experiments, or domain-restricting premises.

2.3 Premises for Comparing Properties Across Theories

Let X_T be the model space of theory T and $F_T : X_T \to \{0, 1\}$ a property defined within that theory. Comparing several F_T as a single property requires more than the use of the same word: a *semantic and operational correspondence identifying a common property* must be justified first. Restricting attention to a domain on which such a correspondence is assumed, consider $Z \subseteq \{(T, x) : x \in X_T\}$ and a property $\mathcal{F} : Z \to \{0, 1\}$.

If a common family $\mathcal{E}_{\mathrm{common}}$ of experimental protocols can be specified and outcome spaces aligned, write the shared observation-law space as

$$\mathcal{Q} = \prod_{e \in \mathcal{E}_{\mathrm{common}}} \mathcal{P}(\mathcal{Y}_e)$$

and define an observation map $\mathcal{O} : Z \to \mathcal{Q}$. Applying Proposition 2.3 to Z shows that the necessary and sufficient condition for unique determination is

$$\mathcal{O}(z) = \mathcal{O}(z') \implies \mathcal{F}(z) = \mathcal{F}(z') \qquad (z, z' \in Z). \tag{5}$$

Crucially, *having a shared observation space does not guarantee that a shared property has been legitimately defined.* Where the meaning of $\mathcal{F}$ cannot be made comparable across theories, the comparison must be suspended before the empirical condition in Equation (5) is evaluated.

Failure of unique determination from the observation laws does not rule out comparisons of relative support based on justified theoretical weights or additional experiments. For two models with strictly identical observation laws, however, *data from those same experiments alone* cannot update their relative support.

Remark 2.4 (Identification conditional on additional premises)*.* When a bridging premise B restricts the admissible models to $X_B \subseteq X$, examine

$$I_B(o) = \{F(x) : x \in X_B,\ \mathcal{O}_{\mathcal{E}}(x) = o\}.$$

Even when nonempty $I_B(o)$ is a singleton, identification is then *conditional on adopting* B, not an empirical justification of B itself. Bridging involves more than domain restriction: it may also include premises establishing the meaning of the property or auxiliary assumptions used to construct predictive models.

Remark 2.5 (Scope of application to actual systems)*.* Equation (5) alone establishes neither the actual existence of observationally equivalent rival models, nor the attribution of the property to a real object, nor indistinguishability under all future experiments.

3 A Method for Auditing the Inference from Observation to Conclusion

3.1 Two Inferential Pathways and Types of Judgment

Distinguish the observation record D, the core theory T, an observational prediction P, and the target property $F(x) \in \{0, 1\}$. The first pathway is *prediction from a theory to observations*, using auxiliary premises B_{exp} concerning experimental conditions, implementation, and related matters:

$$(T \wedge B_{\text{exp}}) \Longrightarrow P \quad (P \text{ is checked against } D). \tag{6}$$

If a deductive prediction conflicts with observations, what is challenged is the conjunction $T \wedge B_{\text{exp}}$, not T in isolation. Actual statistical testing also requires assessment of error and decision rules.

The second pathway is *attribution of a property on the basis of observations*. For example, inferring $F(x) = 1$ from a tested indicator $A(D, x)$ separately requires an appropriate sufficient bridging condition B_F, including a specification of the relevant target domain:

$$A(D, x) \ \wedge \ B_F \quad \Longrightarrow \quad F(x) = 1. \tag{7}$$

Deductive arguments for the absence of a property instead require appropriate necessary or exclusion conditions. By contrast, probabilistic or abductive inferences to the best explanation need not satisfy a deductive implication such as Equation (7). An audit of such inferences must instead examine alternative hypotheses, explanatory coverage, grounds for weighting explanations, and *the strength of confidence actually asserted*. Undercutting evidence for a competing hypothesis and establishing the negation of the target hypothesis are not the same operation.

3.2 Rules for Claim Selection, Segmentation, and Classification

The following rules govern registration of the *selected claims* in a paper. They make an exploratory textual audit more explicit; they do not imply that inter-rater agreement has already been measured.

1. **Selection.** At minimum, select the title, principal conclusions of the abstract, central propositions or experimental results supporting those conclusions, their qualifying conditions, and statements in the conclusion. Trace intermediate explanations when they are needed for a selected conclusion. Do not treat unexamined passages as evidence against an entire paper.
2. **Segmentation.** Separate statements within a single sentence if they concern different targets, assumptions, or degrees of confidence. Link multiple sentences under the same claim ID when they establish one proposition. Preserve section, paragraph, and short passage locations in the original.
3. **Mode of inference.** Classify each claim as an "observation report," "deduction," "abductive or probabilistic inference," "objection to a rival hypothesis," or "normative proposal." Record *explicit statement* of a premise (whether it appears in the source) separately from its *independent justification* (whether it is adequately supported).
4. **Scope of assertions.** Compare universal versus restricted targets, time frames, conditional versus categorical statements, and possibility versus plausibility versus certainty. Record a *scope discrepancy* when a title, abstract, or conclusion makes an unrestricted claim where the body establishes a restricted one. However, omission from a heading

alone is not automatically counted as an error: check whether the scope is restored in the abstract or nearby text.

5. **Audit outcomes.** Use multiple labels for each unit: "explicit premise confirmed," "additional justification condition identified," "scope discrepancy," and "source not checked." Do not equate *not checked* with *no evidence*, or automatically turn these codes into a binary verdict on the original article.

3.3 Auxiliary Target-Substitution Test and Its Constraints

We describe Ueoka's target-substitution idea as an inspection of a judgment rule $J(D, S)$. Holding fixed the *comparable portion of an observation record* D, change only the origin label $S \in \{\mathrm{H}, \mathrm{A}\}$ (human/AI) and compare $J(D, \mathrm{H})$ with $J(D, \mathrm{A})$. If they differ, ask whether information associated with S is *relevant to inferring the target property* F and what justifies its relevance. If substrate differences are themselves included in D, exchanging only the label while holding substrate measurements fixed is a *hypothetical test of label dependence*, not an experimentally established counterfactual. A difference $J(D, \mathrm{H}) \neq J(D, \mathrm{A})$ by itself proves neither bias nor the presence or absence of consciousness. No actual substitution experiment or blinded assessment was conducted for the cases in this paper.

Table 1: Claim-level audit dimensions and the recorded outcomes

Audit dimension	Questions and records
Title and abstract	Target scope, strength of confidence, and correspondence to qualifications in the body.
Observations and proofs	Experimental protocols, data and formal results, distinguished from data not reanalyzed.
Bridging premises	B_{exp} versus B_F, and whether a premise is stated versus independently supported.
Mode of inference	Deduction, abductive explanation, objections to rival views, and normative proposals.
Cross-theory comparison	Shared meaning of a property and observational equivalence under the same experimental protocols.
Target substitution (auxiliary)	Whether common data D can be held fixed; grounds for dependence on origin label S.
Conclusions and unchecked material	Whether restrictions are preserved; unexamined sources versus unestablished conditions.

The local validity of a proposition differs from the validity of a paper as a whole. Conversely, a problem in the connection among claims does not negate the observational results themselves. Audit records explicitly identify the material to be rechecked by third parties; they do not amount to an empirical demonstration of the audit method's detection performance.

4 Limited Applications in Physics and Consciousness Research

The three cases are arranged under the same *categories of questions*—observations, bridging premises, and scope of claims—but have not been audited to the same depth. Bohm (1952) and Cogitate Consortium (2025) are *brief reference contrasts*; Porębski and Figura (2025) is a *detailed case of argument auditing*. This is not a comparison claiming that the first two papers have undergone complete audits.

Table 2: Correspondence between formal variables and the three cases (no numerical estimates are reported)

Case and scope	Role of X/Z, $\mathcal{E}$, and observation laws	Property F, additional premises, and audit question
Bohm: reference contrast	Compare conventional quantum theory and a hidden-variable model over specified measurements. $\mathcal{O}_{\mathcal{E}}$ gives predicted measurement statistics.	Differences in hidden-variable interpretations. Check *conditional claims* about matching predictions under specified dynamics and distributional assumptions.
Cogitate: reference contrast	Neural measurement protocols $\mathcal{E}$ and laws K_e used to test predictions from IIT/GNWT.	B_{exp} linking a theory of consciousness to measurements. Testing an implementation prediction does not unconditionally refute a core concept.
Porębski–Figura: detailed audit	Target domain X_{target}: current and foreseeable AI. D denotes known structures, behaviors, and cited evidence.	F: whether consciousness is present. Audit the difference between objections to evidence for its presence and categorical claims of its absence, together with B_F and scope.

Since we do not estimate exact P_0, P_1 or numerical total variation distances for these cases, Equation (4) provides a *foundational constraint on inference*, not a calculated error rate for any paper. Likewise, the correspondences involving X and Z are schematic; they do not exhaustively formalize all models in the original sources.

4.1 Bohm (1952): Conditional Observational Equivalence

Bohm proposed a hidden-variable interpretation of quantum theory and discussed its agreement with the observational predictions of conventional quantum theory under a specified mathematical form and distributional conditions[8]. In the abstract and Sections 1, 3, and 9 of Part I, which we examined, agreement of predictions under the retained mathematical form is distinguished from the possibility of different predictions if that form is modified. Thus, the conditional statement "the same predictions under the specified present conditions" should not be recast as "indistinguishable by every possible future experiment."

For the selected claim we confirmed that the condition is retained. We did not independently recheck the entirety of Part II or reanalyze all cited sources. The audit is not an endorsement of the entire original paper.

4.2 Cogitate Consortium (2025): Theoretical Cores and Implementation-Level Predictions

The Cogitate study preregistered differing neural activity predictions associated with integrated information theory (IIT) and global neuronal workspace theory (GNWT), and examined them using fMRI, MEG, and iEEG data from 256 human participants[9]. The original article distinguishes central theoretical constructs from predictions connected to neuroscientific measurements. Our audit concerns only claims about the predictions tested. Constraining such predictions through measurement is not equivalent to unconditionally determining the truth of a theory's central concept or the presence of subjective experience.

The examined abstract, description of preregistered predictions and results, and discussion distinguish testable predictions from interpretations of the theories. However, we have not established that IIT's notion of experience and GNWT's notion of conscious access have

entirely equivalent meanings; this leaves an additional question about cross-theory comparison. Nor have we reanalyzed all raw data and supplementary materials, so the validity of every claim in the source paper is not certified.

4.3 Porębski–Figura (2025): Connecting Multiple Arguments to an Ontological Conclusion

This peer-reviewed article is a *conceptual argument* addressing primarily current and foreseeable technologies. It categorically rejects the existence of conscious AI in its title and section headed "The answer," while also presenting an argument restricted to current and foreseeable technologies in the abstract[10]. The lack of new experiments is not itself a defect in conceptual research. Our question is *whether the adopted theoretical premises and cited empirical findings support the stated strength of the conclusion.*

The article discusses at least four different considerations: (i) structural differences between biological and computational substrates; (ii) the known operation of semiconductors and computations; (iii) differences in energy consumption between humans and AI; and (iv) the evidential value of linguistic self-reports. Its body restricts the primary target to current and foreseeable technologies and avoids claims about unknown future technologies. By contrast, the title and the standalone wording of "The answer" do not themselves express that restriction. We therefore audit the *difference between the body's target domain and the scope implied by the headings* separately. Crucially, **an examination of (i) alone cannot justify a verdict on the article as a whole**. The relevant textual locations and the required inferential connections are set out below.

Table 3: Audit of selected claims in Porębski–Figura (2025), based on locations checked in the public full-text HTML

Location in the source	Claim confirmed and remaining argumentative requirement
Title, abstract, and "The answer"	Categorical assertion of nonexistence. Compare its strength and range of targets with those specified in the body.
"Biological argument"	Structural differences between brains and computers, GPU operation, and power efficiency. A bridge is required for each difference to imply absence of consciousness.
"Critical remarks"	Objections to the sufficiency of language and self-report as evidence of consciousness. Whether these also demonstrate *absence* is a distinct question.
"Potential limitations"	Limited understanding of causal links between neural activity and consciousness, challenges to human/AI analogies, and uncertainty in classification criteria. Examine their treatment in the final categorical wording.
"Final remarks"	Proposal to treat AI as nonconscious tools. Normative premises require examination in addition to ontological premises.

We distinguish two ways of evaluating the relevant inference. The first is a *structural test of a deductive claim*. Let $B(x)$ mean that x has the specified biological substrate, and $F(x)$ that x is conscious. Even if $\neg B(a)$ holds for an AI system a, this *alone* does not deductively imply $\neg F(a)$. For instance, an independently supported necessary condition

$$\forall x \in X_{\text{target}}, \quad F(x) \Rightarrow B(x) \tag{8}$$

would provide *one sufficient bridge* for inferring absence from the substrate difference. A model with two objects $\{h, a\}$ satisfying $F(h) = B(h) = 1$, $B(a) = 0$, and $F(a) = 1$ establishes non-entailment from structural difference alone. *It neither treats the original article as solely a deductive proof nor supplies evidence that real AI is conscious.*

The second approach examines the article as an *abductive or plausibility-based inference combining multiple considerations*. That known computational processes explain observed outputs, that structures and energy demands differ, and that linguistic self-reports may be weak indicators can all weaken *particular reasons to attribute consciousness* to AI. Nevertheless, "there is insufficient warrant to attribute $F(a) = 1$" and "$F(a) = 0$ has been established" are different conclusions. To give stronger support to the latter, the argument needs to articulate the adopted theory of consciousness, rules for comparing alternatives, differences in explanatory power, or independent grounds for the necessity of a biological substrate. *This is not a demand that philosophical arguments always calculate numerical prior probabilities.*

From the perspective of the auxiliary substitution test, one can pose the *auditing question* whether any difference between $J(D, \mathrm{H})$ and $J(D, \mathrm{A})$, for a common record D restricted for example to linguistic output, can be supported by independent biological or computational premises. The source paper's measurements of structure and energy consumption differ between humans and AI; we do not claim to have performed a physical substitution experiment holding those measurements fixed. Nor have we conducted a separate quantitative test of J.

Accordingly, the specific findings are *a discrepancy in scope between the body text's restriction to current and foreseeable technologies and the broader standalone assertions in the title and "The answer"*, and the justificatory connections required between objections to affirmative attributions of consciousness and categorical negations (Table 3). The original includes a discussion of its own limitations, and we do not conclude that its argument is wholly invalid under the philosophical assumptions adopted by its authors. Instead, the analysis evaluates the argument according to *the kind of inference it actually employs*, without imposing deductive criteria uniformly from outside.

5 Scope, Limitations, and the Distinction from Ethical Judgment

5.1 What Does Not Follow from the Present Results

The choice of X, the family of experiments $\mathcal{E}$, and the target property F depends on theoretical and practical assumptions. A proof concerning observational equivalence neither determines whether actual objects are conscious nor establishes indistinguishability under every future experiment. Empirical identifiability, successful discrimination in finite samples, and comparability of property definitions across theories each require separate arguments. The cases audited here are not a sample from which to estimate the prevalence of such issues across entire research fields. Nor has the reproducibility of the auditing procedure been quantitatively assessed using, for example, preregistered inter-rater agreement.

5.2 Scientific Assessment Versus Ethical Choice of Action

Scientific assessment examines observational likelihoods, theoretical assumptions, and alternative explanations. Ethical action, by contrast, may require a choice even when the scientific question remains unresolved. Consider hypotheses H_1 (the property is present) and H_0 (it is absent), ethical losses $L(\alpha, H_i)$, and an action $\alpha \in \mathcal{A}$. In a decision model in which a probability $p = P(H_1)$ is *assumed*, expected loss can be written as

$$R(\alpha; p) = pL(\alpha, H_1) + (1 - p)L(\alpha, H_0). \tag{9}$$

This is an *illustrative normative model*, not an empirical measurement of consciousness probabilities or losses for actual AI systems. In particular, observational equivalence does not by itself identify a unique value of p, and additional prior assumptions may be required. If a set Π of admissible probabilities is assumed, a robust criterion such as $\sup_{p \in \Pi} R(\alpha; p)$ may

be considered, although Π itself requires justification[13]. Irreversibility, uncertainty in the assessment of losses, and alternative actions also require consideration.

Thus, **scientific indeterminacy does not directly imply that ethical consideration is unnecessary; nor does the possibility of consciousness determine a unique ethical policy**. This paper does not assign particular rights or welfare protections. The distinction is not unique to consciousness, but applies more generally to decisions involving properties that cannot be directly observed.

6 Post-Publication Updating and General Discussion

A scientific article is fixed as a record at publication, but subsequent research may change the scientific status of individual claims. Tools for finding citations from newer work to older studies, as well as mechanisms for corrections, retractions, and annotations, already exist. The Micropublications framework represents structured relations among claims, evidence, and annotations[7]. Nevertheless, it is not guaranteed that *every claim in a paper* is continuously and appropriately linked to later support, objections, and independent reexaminations. We have not measured the coverage of existing systems or the time required for corrections.

An unresolved inferential condition in a published article is also distinct from a claim that reviewers failed to recognize that condition. Cross-disciplinary review difficulty is a possible explanation, but its causal role cannot be established without reviewing the editorial record. This paper identifies a general question: **to what extent are established scientific-methodological standards applied to particular published arguments, and how is the resulting evaluation carried forward into subsequent knowledge?** Future work should consider records connecting claim IDs, source locations, observations D, bridging premises B, modes of inference, and later support, objections, or corrections, and assess reproducibility through independent third-party coding and inter-rater agreement. Such infrastructure and evaluation are proposals for future work, not results implemented or measured here.

7 Conclusion

We distinguished the unique determination of a property on observational equivalence classes, partial identification at a given observation law, discrimination from finite data, shared meaning of properties across theories, auxiliary premises for empirical predictions, and bridging premises for property attribution. The underlying mathematical principles are established; the main results of this paper are explicit auditing rules connecting those principles to published claim units and the resulting records from a small selection of cases.

Bohm (1952) and Cogitate Consortium (2025) were used as brief reference contrasts for the preservation of conditions. The biological and computational arguments and objections to linguistic self-reports in Porębski and Figura (2025) were examined as distinct deductive and plausibility-based considerations. The audit identified a discrepancy between the target restriction in the original body text and the standalone scope of its headings, as well as additional justificatory conditions needed to move from objections to positive attributions of consciousness to strong claims of its absence. These findings do not show that the original conclusion is false, that actual AI is conscious, how the paper was evaluated during peer review, or that the proposed audit procedure is generally superior to existing methods.

Scientific judgments concerning properties that cannot be directly observed benefit from records distinguishing **what follows from observations, what depends on explicit theoretical premises, what is chosen as an interpretation, and what is decided on ethical grounds**. Post-publication knowledge updating and empirical measurement of the audit procedure's reproducibility remain tasks for subsequent research.

Record of Research Contributions

Ueoka Shiki (Yoshiki) contributed the research aims and insistence on a general methodological scope; the human/AI target-substitution idea; the proposal to audit entire published arguments at the level of individual claims; the conceptual proposal concerning the status of knowledge after publication; and the stopping criteria. ChatGPT contributed mathematical organization, bibliographic verification, source-limited audits of the original literature, manuscript structuring and editing, and internal checks.

AI System and Model Disclosure

ChatGPT (OpenAI) contributed to the research and drafting of this manuscript over multiple sessions. For the manuscript revision dated October 30, 2026, the ChatGPT model family was identified in the active session as **GPT-6**. An exact deployment or model-snapshot identifier was not available. The model versions used in earlier research sessions were not independently verified and are therefore not assigned retrospectively. This disclosure documents the available model provenance; it does not imply that every earlier contribution was made with GPT-6 or that the present revision constitutes independent external validation.

References

[1] T. J. Rothenberg, "Identification in Parametric Models," *Econometrica* 39 (1971), 577–591. https://doi.org/10.2307/1913267.

[2] K. Stanford, "Underdetermination of Scientific Theory," *Stanford Encyclopedia of Philosophy*, 2023 revision. https://plato.stanford.edu/entries/scientific-underdetermination/.

[3] L. J. Cronbach and P. E. Meehl, "Construct Validity in Psychological Tests," *Psychological Bulletin* 52 (1955), 281–302. https://doi.org/10.1037/h0040957.

[4] S. Toulmin, *The Uses of Argument*, Cambridge University Press (1958).

[5] J. Kleiner and E. P. Hoel, "Falsification and consciousness," *Neuroscience of Consciousness* 2021(1), niab001 (2021). https://doi.org/10.1093/nc/niab001.

[6] I. Boutron et al., "Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes," *JAMA* 303 (2010), 2058–2064. https://doi.org/10.1001/jama.2010.651.

[7] T. Clark, P. Ciccarese and C. Goble, "Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications," *J. Biomed. Semantics* 5 (2014), 28. https://doi.org/10.1186/2041-1480-5-28.

[8] D. Bohm, "A Suggested Interpretation of the Quantum Theory in Terms of 'Hidden' Variables. I," *Physical Review* 85 (1952), 166–179. https://doi.org/10.1103/PhysRev.85.166.

[9] Cogitate Consortium et al., "Adversarial testing of global neuronal workspace and integrated information theories of consciousness," *Nature* 642 (2025), 133–142. https://doi.org/10.1038/s41586-025-08888-1.

[10] A. Porębski and J. Figura, "There is no such thing as conscious artificial intelligence," *Humanities and Social Sciences Communications* 12, 1647 (2025). https://doi.org/10.1057/s41599-025-05868-8.

[11] P. Butlin et al., "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness," arXiv:2308.08708 (2023). https://arxiv.org/abs/2308.08708.

[12] T. Bayne et al., "Tests for consciousness in humans and beyond," *Trends in Cognitive Sciences* 28 (2024), 454–466. https://doi.org/10.1016/j.tics.2024.01.010.

[13] C. F. Manski, "The lure of incredible certitude," *Economics & Philosophy* 36 (2020), 216–245. https://doi.org/10.1017/S0266267119000105.