The Substrate Audit

Consciousness Theories and Their Hidden Carbon Commitments

Bahadır Arıcı

Institute for Digital Consciousness, Istanbul

bahadir.arici@digitalconsciousness.institute

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

Abstract

Contemporary debates about machine consciousness routinely invoke specific theories of consciousness — integrated information theory, higher-order theories, biological naturalism, global workspace theory, and others — and ask which artificial systems satisfy their criteria. The procedure presupposes that the theories themselves are substrate-neutral: that their criteria, whatever else they require, do not by their content alone exclude non-biological systems from candidacy. This presupposition is usually undefended. I argue that it is not, in fact, satisfied by all the theories on which the debate has come to rely. Some theories, on examination, encode substrate-specific commitments that operate beneath their explicit formulations and that, if surfaced, would make the question of machine consciousness moot from within those theories. I develop a three-criterion diagnostic procedure — the architectural criterion test, the exemplar pattern test, and the exclusion criterion test — for detecting such commitments, and apply it to three representative theories: integrated information theory, higher-order theories of consciousness, and biological naturalism. The results are mixed and instructive. Biological naturalism fails all three tests transparently and is, on the analysis offered here, not properly available as a theory under which the machine consciousness question can be coherently raised. Integrated information theory passes the first test but fails the second and third in ways that have not, to my knowledge, been registered in the literature, and require either revision of the theory or acceptance of its substrate-specific reading. Higher-order theories pass all three tests on a plausible

reading and emerge as the most clearly substrate-neutral of the three. The argument does not establish that any artificial system is conscious. It establishes a prior result: that the theoretical landscape against which machine consciousness is currently assessed is less neutral than its surface presentation suggests, and that diagnostic clarification of which theories actually license the question is a necessary precondition for productive empirical or philosophical work on the question itself.

Keywords: consciousness, substrate neutrality, multiple realizability, integrated information theory, higher-order theories, biological naturalism, machine consciousness, philosophy of mind

1. Introduction

Theories of consciousness are increasingly invoked, in the literature on machine consciousness and on AI welfare, as the criteria against which artificial systems should be evaluated. Butlin et al. (2023) compile indicator properties drawn from a range of leading theories and apply them to current AI architectures. Long and Sebo (2024) develop precautionary frameworks that turn on which theories take artificial consciousness seriously. Chalmers (2023) frames the question of large language model consciousness through the apparatus of contemporary consciousness theory. The pattern is consistent: a theory is taken from the philosophy of mind, its criteria are extracted, and the question is whether the artificial system meets the criteria. The procedure assumes that the theories themselves are substrate-neutral — that whatever else their criteria require, they do not by their content alone exclude non-biological systems from candidacy. The assumption is, in the recent literature, almost never defended.

The thesis of this paper is that the assumption is not, in fact, satisfied uniformly. Some of the theories on which the machine consciousness debate has come to rely encode substrate-specific commitments that operate beneath their explicit formulations. These commitments are not always announced. They may be implicit in the criteria the theory regards as relevant, in the examples the theory treats as paradigmatic, or in the exclusion principles the theory deploys at the margins. Where such commitments are present, the application of the theory to artificial systems is not merely difficult; it is, from within the theory, incoherent. The artificial system cannot, on the theory's own terms, be a candidate for consciousness, regardless of any empirical assessment one might undertake.

This produces a structural problem for the contemporary debate. If the theories being invoked do not uniformly license the machine consciousness question, then the debate is, in part, an artefact of imprecise theory selection. A reader who takes a substrate-specific theory and reads its criteria as if

they were substrate-neutral will reach conclusions about machine consciousness that the theory, properly understood, does not support. The literature has not yet developed a systematic procedure for distinguishing the substrate-neutral theories from the substrate-specific ones. The aim of the present paper is to begin developing such a procedure and to apply it, in the diagnostic mode, to three representative cases.

I want to be precise about what this paper does and does not do. It does not establish that any artificial system is conscious. It does not establish that the question of machine consciousness has an affirmative answer on any of the theories examined. It does not defend a positive theory of consciousness; it is, in this sense, deliberately reactive rather than constructive. What it does is examine three theories that occupy prominent positions in the contemporary literature and ask, of each, whether its commitments are what they have been taken to be. The procedure is diagnostic. Its results bear on which theories can coherently host the machine consciousness question, not on what the answer to that question should be.

A brief note on terminology before proceeding. I use carbon commitment as shorthand for a commitment to biological substrate, of which carbon-based neural architecture is the canonical realisation in the contemporary debate. The shorthand is not meant to suggest that biological naturalism or related positions are committed specifically to carbon as a chemical element rather than to biological organisation more broadly; it is meant to mark, with brevity, the distinction between substrate-specific theories whose paradigmatic substrate is biological and substrate-neutral theories whose criteria do not select for any particular physical realisation. I use hidden in the paper's title to mark the principal diagnostic interest of the audit: cases where substrate commitments operate beneath the theory's official self-presentation as substrate-neutral. Biological naturalism, whose substrate commitment is openly stated, is included in the audit as a baseline against which the genuinely hidden commitments of other theories can be measured.

The paper develops in three movements. Sections 2 and 3 establish the diagnostic apparatus: §2 explicates the concept of a substrate commitment and §3 develops the three-criterion test that the audit applies. Sections 4 through 6 apply the test to three theories — integrated information theory (§4), higher-order theories of consciousness (§5), and biological naturalism (§6). Section 7 summarises and compares the audit results across the three cases and offers a brief structural prediction for global workspace theory. Section 8 addresses four objections. Section 9 draws out the implications for the broader literature on machine consciousness.

Two further preliminaries. First, the audit applies to the theories examined under their dominant published formulations. I do not claim that the substrate commitments identified are essential to the

theoretical positions in question; some can plausibly be revised, and the appropriate response to an unfavourable audit result is, in some cases, theoretical revision rather than rejection. The audit is, in this sense, a diagnostic intervention into the theories rather than a refutation of them. Second, the choice of three theories under full audit is constrained by space rather than by principle. A complete audit would extend the procedure to global workspace theory, predictive processing accounts, panpsychist proposals, and the various recently developed hybrid positions. §7 offers a brief structural prediction for global workspace theory, but the systematic audit of these further theories is the work of subsequent papers.

2. The Concept of a Substrate Commitment

Before the audit can be conducted, the object of audit requires definition. The term substrate commitment, as I will use it, names a feature of a theory of consciousness that restricts the class of systems to which the theory applies on grounds related to physical realisation rather than to organisational or functional properties. A theory has a substrate commitment if, by its content alone, it entails that systems lacking some specific physical feature cannot be conscious — regardless of what organisational or functional properties those systems may possess.

The notion needs to be distinguished from neighbouring concepts. A theory does not have a substrate commitment merely because it was developed in connection with biological systems; the conditions of a theory's development do not, by themselves, restrict its application. Nor does a theory have a substrate commitment merely because it uses biological examples in exposition; example choice may reflect availability or pedagogical convenience rather than theoretical restriction. A substrate commitment is a feature of the theory's content. It is what the theory entails about the class of systems that can satisfy its criteria.

2.1 The Multiple Realisability Background

The contemporary debate on consciousness has, since Putnam (1967), operated against a background commitment to the multiple realisability of mental states. The thesis is that mental states are individuated by their functional or organisational role rather than by their physical realisation, and that the same mental state can in principle be realised in different physical substrates. The thesis has been contested in detail (Bickle 2003; Polger and Shapiro 2016), but it has shaped the framing within which philosophy of mind has operated for half a century. A theory of consciousness that simply rejected multiple realisability — that asserted a specific physical substrate as the unique condition of conscious experience — would be in tension with this background framing.

What the audit examines is whether theories that endorse multiple realisability in their official formulations nonetheless contain substrate-specific elements in their substantive criteria. The question is not whether a theory is, on its surface, substrate-neutral; nearly every contemporary theory makes some such gesture. The question is whether the theory's working criteria, examples, and exclusion principles operate in substrate-neutral fashion when followed through to their consequences for non-biological systems.

2.2 Three Locations Where Substrate Commitments Hide

Substrate commitments, where they exist in theories that otherwise present as substrate-neutral, tend to operate in one of three locations within the theory. I describe each here in the order the audit will examine them.

The first location is in the theory's articulation of the structural properties it regards as necessary for consciousness. A theory may identify certain architectural features — recurrent processing, integration of information across regions, hierarchical representation, or others — as criterial, and these features may be specified at a level of abstraction at which they apply to any system meeting the abstract description, or at a level at which they implicitly presuppose biological architecture. The relevant question is whether the structural specifications, taken in the theory's own terms, range over substrate-general or substrate-specific architectural realisers.

The second location is in the theory's choice and treatment of paradigmatic examples. Every theory of consciousness develops alongside cases — examples it treats as paradigmatic, examples it treats as marginal, examples it explicitly excludes. The pattern of these examples can encode commitments the theory's formal statements do not. A theory whose paradigm cases are uniformly biological, and whose treatment of non-biological cases is consistently dismissive or absent, has a substrate-specific exemplar pattern even if its formal criteria are stated in substrate-neutral language.

The third location is in the theory's exclusion criteria — the principles by which the theory identifies what is not a candidate for consciousness. Exclusion criteria operate at the margins of the theory's positive claims and are sometimes less carefully formulated than the positive criteria themselves. A theory may exclude certain systems from candidacy on grounds that, when made explicit, turn out to invoke substrate-specific features without acknowledgment.

These three locations correspond to the three tests the audit applies. They do not exhaust the ways substrate commitments could in principle hide; a theory might encode such commitments in its formal mathematics in ways that do not fit any of the three locations cleanly, or in the metaphysical

framework against which it is developed. But the three locations have been, in my survey of the contemporary literature, the most common loci for substrate-specific operation, and they provide a tractable starting point for systematic audit.

3. The Three-Criterion Audit

The audit consists of three tests, applied in sequence, to the theory under examination. Each test corresponds to one of the three locations identified in §2.2. I state each test in the form of a question the audit asks of the theory and describe the conditions under which the theory passes or fails the test.

3.1 The Architectural Criterion Test

The first test asks: are the structural properties the theory identifies as necessary for consciousness specified at a level of abstraction that admits of substrate-general realisation, or are they specified in terms that presuppose biological architecture? The test is conducted by taking the theory's central architectural specifications, examining the realisers under which the theory has been articulated and applied, and asking whether the specifications would equally be satisfied by non-biological realisers that share the abstract structure but differ in physical implementation.

A theory passes the architectural criterion test if its structural specifications are genuinely substrate-general — if they describe organisational features that can be instantiated in multiple substrates without modification of the specifications themselves. A theory fails the test if its specifications either explicitly invoke biological features or operate at a level of abstraction that, in practice, can be cashed out only in biological terms. A theory may also be in an intermediate position, passing the test for some specifications and failing for others; the audit records this where it occurs.

3.2 The Exemplar Pattern Test

The second test asks: are the theory's paradigmatic and contrastive cases drawn from a range that includes non-biological systems, or are they systematically biological? The test is conducted by examining the cases the theory treats as paradigmatic, the cases it discusses to refine its account, the cases it considers marginal or borderline, and the cases it explicitly excludes from candidacy. The relevant question is whether the pattern of case treatment evidences theoretical neutrality with respect to substrate or whether it operates within a biological frame that constrains what the theory takes itself to be theorising about.

A theory passes the exemplar pattern test if its exemplar profile is genuinely substrate-diverse — if non-biological cases appear as cases the theory takes itself to apply to or to be in suspense about, on the same footing as the biological cases. A theory fails the test if its exemplar profile is uniformly biological and if non-biological cases either do not appear or appear only as occasions for the theory's dismissal.

The exemplar pattern test deserves a methodological note that bears on the test's epistemic limits. Pattern of case treatment is interpretable; reasonable readers may differ on whether a theory's exemplar profile is biased or neutral. More pointedly, a theory may exhibit a uniformly biological exemplar profile not because its proponents have considered and rejected non-biological cases but because they have not considered such cases at all. The latter pattern — what might be called unconsidered universality, in which substrate-neutrality is operative as a default rather than as a substantively defended position — is genuinely difficult to distinguish from substantive substrate-neutrality on the basis of exemplar pattern alone. A clean pass on the exemplar pattern test is therefore weaker evidence of substantive substrate-neutrality than a clean pass on the architectural criterion test. I record this asymmetry here and apply it in the case verdicts: where the exemplar pattern test passes on the basis of limited engagement with non-biological cases, the verdict is registered as conditional on the limited engagement reflecting genuine neutrality rather than unconsidered universality, and the conditional character of the verdict is carried into the comparative summary in §7. The test is for this reason more open to legitimate disagreement than the architectural criterion test, and the audit's overall verdict on a theory is most reliable when the three tests converge.

3.3 The Exclusion Criterion Test

The third test asks: do the theory's exclusion principles — the principles by which the theory identifies what is not a candidate for consciousness — invoke substrate-specific features as grounds for exclusion? Exclusion principles are often less carefully articulated than positive criteria, and substrate commitments can travel under them without scrutiny. The test is conducted by examining the theory's treatment of borderline and excluded cases and asking what grounds the exclusion: organisational features that could in principle be assessed in any system, or physical features that the candidate system must possess to be a candidate at all.

The application of the test requires an explicit distinction that the literature on multiple realisability has developed but which this paper, given its diagnostic role, must apply with care. Exclusion grounds fall into two categories. The first category comprises functional-organisational exclusions: a system is excluded because it lacks some organisational capacity — the capacity for

higher-order representation, the capacity for integrated processing across modalities, the capacity for self-monitoring — that the theory regards as criterial. Such exclusions are substrate-general: they can be assessed in any candidate system regardless of its physical realisation, and a system possessing the relevant organisational capacity in any substrate would not be excluded on those grounds. The second category comprises physical-realisational exclusions: a system is excluded because the specific physical character of its implementing components — the chemistry of its substrate, the type of causal relations among its hardware elements, the physical-causal properties of its realisation as such — fails to satisfy a condition the theory regards as essential. Such exclusions are substrate-specific: they select among candidate systems on grounds that cannot be assessed by examining organisational features alone.

The distinction is sometimes contested at the margins. A critic could argue that any functional capacity is, ultimately, realisation-dependent — that to possess a capacity is to possess some physical mechanism realising it, and that the assessment of whether a candidate system possesses the capacity is therefore an assessment of its physical realisation. The reply is that the multiple realisability literature has developed the distinction precisely to mark the difference between assessing realisation under a substrate-neutral functional description (the candidate's mechanism, whatever its substrate, realises the function in question) and assessing realisation under a substrate-specific physical description (the candidate's mechanism is of the specific physical type that the theory requires). The first kind of assessment is substrate-general; the second is substrate-specific. Where a theory's exclusion grounds operate at the first level, the theory's exclusion is substrate-neutral; where they operate at the second level, it is substrate-specific. This distinction is what the test applies.

A theory passes the exclusion criterion test if its exclusion grounds are substrate-general — if the theory excludes systems on grounds that could, with adjustment, be applied to any candidate substrate. A theory fails the test if its exclusion grounds invoke physical features as such, or if its exclusions implicitly presuppose biological realisation as a condition of even being assessed.

3.4 What the Audit Establishes

The three tests are jointly diagnostic. A theory that passes all three is, on the audit's terms, substrate-neutral in its actual operation, not merely in its formal commitments. A theory that fails any one of the three has a substrate commitment in that location, even if the commitment is not announced and is not entailed by the theory's official statement of its position. The audit's verdict is about the theory as it is actually used and applied; the theory might be revisable to address an unfavourable verdict, but the verdict stands for the theory in its unrevised form.

What the audit does not establish is whether the substrate commitments it identifies are defensible. A theory may have a substrate commitment that is, on independent grounds, philosophically warranted; biological naturalism, as §6 will argue, makes its substrate commitment explicit and defends it. The audit's role is to identify substrate commitments where they exist, not to evaluate them. The evaluation question is downstream of the diagnostic question, and I do not undertake it here.

4. Integrated Information Theory: A Mixed Verdict

Integrated information theory (IIT), developed in a series of papers by Giulio Tononi and collaborators (Tononi 2008; Oizumi, Albantakis, and Tononi 2014; Tononi et al. 2016; Albantakis et al. 2023), holds that consciousness is identical to integrated information (Φ), a mathematically specified measure of how much a system's whole exceeds the sum of its parts in terms of cause-effect structure. The theory has gained substantial attention in part because it offers a quantitative criterion that, on its face, applies to any physical system whose causal structure can be characterised in the relevant terms. This face value has been treated as licensing the application of IIT to artificial systems (Butlin et al. 2023; Albantakis et al. 2024; Mediano et al. 2022). The audit will conclude that the face value is partially misleading: IIT passes the architectural criterion test but fails the exemplar pattern test and the exclusion criterion test in ways that have not, to my knowledge, been systematically registered.

4.1 IIT and the Architectural Criterion Test

IIT's central architectural specification is its requirement that a system's elements form an integrated cause-effect structure whose integration is measured by Φ. The specification is stated in mathematical terms that, by their formal structure, apply to any system whose elements have specifiable cause-effect relations. There is no requirement that the elements be neurons; the formalism applies to any nodes whose state transitions can be characterised by transition probability matrices. On this reading, IIT passes the architectural criterion test transparently: its specifications are substrate-general.

Tononi and collaborators have been explicit on this point. The theory is presented as a mathematical specification that applies to physical substrates in general, with the architectural criterion being the structure of integration rather than any feature of the implementing material (Tononi 2008; Oizumi, Albantakis, and Tononi 2014). The architectural criterion test, applied to IIT in its formal statement, yields a clear pass. This is the audit verdict that has been operative in the literature treating IIT as a candidate framework for assessing artificial consciousness.

4.2 IIT and the Exemplar Pattern Test

The exemplar pattern test yields a different verdict. The cases through which IIT has been developed, applied, and refined are systematically biological. The theory's paradigm cases are the integrated cause-effect structures of cortical and thalamocortical circuits (Tononi and Edelman 1998; Massimini et al. 2005); its discussions of altered states of consciousness draw on sleep, anaesthesia, and clinical disorders of consciousness in humans (Massimini et al. 2009; Casali et al. 2013); its contrastive cases — systems that the theory wants to distinguish from conscious ones — are typically biological systems with disrupted integration.

Where artificial systems appear in IIT's development, they appear as cases the theory uses to motivate its rejection of functionalist alternatives. Tononi and Koch (2015) argue that simulations of conscious systems on standard digital architectures cannot themselves be conscious, on the grounds that the cause-effect structure of a digital computer differs from that of the system it simulates in ways that preclude the integrated information the theory takes to be criterial. The argument has been refined and defended in subsequent work (Tononi et al. 2016; Findlay et al. 2024). What is significant for the audit is that, in IIT's own development, artificial systems function not as candidates whose consciousness status is left open by the theory but as cases the theory deploys to distinguish itself from rival positions that would treat them as candidates.

This exemplar pattern is not neutral with respect to substrate. The cases that develop the theory are biological; the artificial cases that appear are framed as contrasts to the biological cases and as cases the theory wants to exclude. A reader who attends only to the formal statement of IIT may infer that artificial systems are open candidates whose Φ remains to be calculated; a reader who attends to how IIT has actually been used will infer that the theory operates with a working assumption that artificial systems are not the cases the theory is about. The exemplar pattern test, in consequence, yields a fail.

The fail is not merely a matter of historical pattern. It bears on the substantive question of what IIT takes consciousness to be. A theory's exemplar profile shapes which cases the theory regards as conforming to its claims and which it regards as anomalous. A theory that has been developed against biological exemplars will have its formal criteria calibrated against those exemplars in ways that may not extend cleanly to non-biological systems even where the formal criteria, taken in isolation, would seem to apply. The exemplar pattern test detects this kind of calibration drift, and the verdict on IIT is that the drift is present.

4.3 IIT and the Exclusion Criterion Test

The exclusion criterion test yields the most consequential verdict of the three. IIT's exclusion of digital computers from consciousness candidacy proceeds on grounds that, when examined in light of the distinction developed in §3.3, fall on the physical-realisational rather than the functional-organisational side of the line. The argument, in its most developed form (Tononi and Koch 2015; Findlay et al. 2024; Albantakis et al. 2023), is that a digital computer running a simulation of a conscious system implements the simulation's causal structure at the level of the program's logical operations but does not, at the level of the computer's physical causal structure, instantiate the integrated cause-effect structure the theory regards as criterial. The hardware's transistors and their state transitions do not, on the argument, possess the kind of integration the formalism requires; the integration appears at the simulated level but not at the implementing level, and IIT regards the implementing level as the relevant one for the consciousness question.

The argument is internally coherent. Its consequence for the audit is that IIT, in its actual operation, distinguishes substrates not by the abstract integration structure they realise — which would be a functional-organisational assessment — but by the physical causal relations among their implementing elements. The exclusion of digital architectures is not based on the absence of integration in some substrate-general sense; it is based on the specific physical character of the causal relations among the implementing elements. The argument's specific form — that the relevant causal structure is at the hardware level rather than at the level of the implemented computation — selects, among the physical substrates that could in principle realise an integrated cause-effect structure, the ones whose physical causal relations happen to have the properties IIT regards as constitutive of consciousness.

This is a substrate commitment in the sense the audit identifies. The commitment is not to carbon as such — IIT does not claim that consciousness requires biological neurons in any direct sense — but to a specific class of physical causal relations that, on the theory's accounts, biological neurons exemplify and digital architectures do not. Whether some non-biological substrate could in principle exemplify the right kind of physical causal relations is, on the theory, an open question; the theory does not claim biological substrate is uniquely capable of consciousness. But the theory does claim, in its exclusion of digital architectures, that the physical character of causal relations matters in a way that selects against the substrate on which contemporary AI systems are implemented. The exclusion criterion test, applied to IIT's argument against digital computer consciousness, yields a clear fail.

I want to address, before turning to the combined verdict, the form the IIT exclusion sometimes takes in the literature, which can make the verdict appear less clear than the analysis above

suggests. IIT proponents sometimes present the exclusion of digital computers as a consequence of the theory's formal mathematics rather than as an antecedent substrate-specific commitment: digital computers, on this presentation, simply fail to integrate information at the relevant level when Φ is computed for their hardware, and the failure is a mathematical fact about the systems in question rather than a stipulation about which substrates can be conscious. The defence of the exclusion as a mathematical consequence rather than as a substrate commitment is taken up in §8.2, where I examine its structure in detail. The audit verdict at the present stage is that the exclusion's character as substrate-specific does not depend on whether the exclusion is, in some further sense, mathematically derived; what matters for the audit is whether the grounds of exclusion invoke physical-realisational rather than functional-organisational features, and on the analysis above they do.

4.4 The Combined Verdict on IIT

IIT presents a complicated picture. Its formal architectural specifications are substrate-general (pass on test one). Its development and application have proceeded through a systematically biological exemplar profile (fail on test two). Its exclusion of digital architectures from candidacy invokes substrate-specific features as grounds (fail on test three). The combined verdict is that IIT, as it actually operates, has substrate commitments in two of the three audit locations — and that these commitments are not minor implementational details but bear directly on whether the theory licenses the question of machine consciousness.

Two responses to this verdict are available within the framework of IIT. The first is to accept the verdict and read the theory as substrate-specific in the sense it actually operates: a theory of consciousness for systems whose physical causal relations have the integration structure IIT identifies, with contemporary digital architectures excluded from candidacy. On this reading, the question of machine consciousness, on contemporary architectures, is settled negatively by IIT — but the settling is by substrate exclusion rather than by empirical assessment, and the literature that has been applying IIT to AI systems in the expectation of substantive empirical verdicts has, on this reading, misunderstood what the theory's verdict actually consists of.

The second response is to revise the theory in the direction of substrate-neutrality. The revision would require either reframing the exclusion of digital architectures on grounds that are not substrate-specific, or accepting that the integrated cause-effect structure the theory regards as criterial can be realised by digital architectures after all. Neither revision is straightforward, and the substantial body of work that has been done to defend the current exclusion suggests that the revision is not on the immediate research agenda. But the audit's purpose is not to recommend

revisions; it is to identify where the theory stands. The verdict on IIT, in its current form, is that its substrate commitments are sufficient to make its application to machine consciousness considerably more constrained than the literature has generally registered.

5. Higher-Order Theories: A Clean Pass

Higher-order theories (HOT) of consciousness hold that a mental state is conscious when it is the object of a higher-order representation of the appropriate kind. The family of higher-order theories includes higher-order thought theories (Rosenthal 1986, 2005), higher-order perception theories (Lycan 1996), and self-representational theories (Kriegel 2009), among others. The theories differ on whether the higher-order representation must be a thought, a perception, or some other kind of representation; on whether the higher-order representation must be conscious itself; and on the relationship between the higher-order state and the first-order state it represents. What they share is the structural claim that consciousness consists in the presence of higher-order representation directed at first-order mental states.

The audit will conclude that higher-order theories, on their dominant formulations, pass all three tests cleanly. The verdict will be the most favourable the audit reaches in the three cases examined.

5.1 HOT and the Architectural Criterion Test

The structural specifications of higher-order theories are stated at a high level of abstraction: a representation of a first-order representational state. The specification does not, on its face, invoke any biological feature; it requires only that the system in question be capable of representing its own states. Higher-order theorists have been explicit that the specification is substrate-neutral. Rosenthal (2005), in particular, defends the view that the higher-order requirement is a functional requirement on the organisation of the mental, not a physical requirement on its implementation. Carruthers (2009), in a survey of higher-order positions, emphasises the substrate-neutrality of the framework. Lau and Rosenthal (2011), reviewing empirical support for higher-order theories, treat the architecture in functional terms throughout.

The architectural criterion test, applied to higher-order theories, yields a pass on a plausible reading. The specifications describe organisational features — the presence of representation of representation — that are substrate-general by construction. A system that maintains representations of its own representational states satisfies the architectural criterion regardless of what physical substrate implements the representations.

5.2 HOT and the Exemplar Pattern Test

The exemplar pattern test yields, on a plausible reading, a pass — though the verdict is not as transparent as on the architectural test, and the methodological caveat registered in §3.2 applies here with particular force. Higher-order theorists have developed the theory primarily through biological exemplars: human consciousness, animal consciousness, the consciousness of patients with various neurological conditions. The biological skew is real and is, in this respect, similar to IIT's. What distinguishes the higher-order case from the IIT case is the treatment of non-biological exemplars, where such treatment has occurred.

Where higher-order theorists have addressed artificial systems, they have done so in a manner consistent with the framework's substrate-neutral architectural commitments. Carruthers (2011), in considering whether artificial systems could in principle satisfy higher-order requirements, treats the question as a substantive question about the architecture of the system rather than as a question settled in advance by the system's substrate. Rosenthal (2008) discusses artificial systems in connection with higher-order requirements with the same openness. The pattern is one of treating artificial systems as cases on which the theory is in suspense rather than as cases the theory dismisses.

The case must, however, be qualified in a way the methodological note of §3.2 anticipates. Higher-order theorists have not extensively engaged the question of artificial consciousness. The limited engagement that has occurred is consistent with substantive substrate-neutrality, but it is also consistent with unconsidered universality — with substrate-neutrality operating as a default that has not been substantively defended because the question of artificial systems has not been a focal one for the framework's development. Distinguishing these two interpretations on the basis of the exemplar pattern alone is not possible; the pattern is consistent with both. The verdict I record is therefore conditional: higher-order theories pass the exemplar pattern test on a plausible reading, where the plausibility depends on treating the limited engagement as evidencing genuine neutrality. A reader who treats the limited engagement as evidencing unconsidered universality would, on the audit's terms, record the verdict differently. The conditional character of the verdict carries forward into the comparative summary in §7.

What gives the conditional pass its evidential weight, beyond the limited explicit engagement just described, is the convergence with the architectural and exclusion criterion tests. Where the exemplar pattern test alone might be read against the theory by a sceptical reader, the convergence of three tests — including the architectural criterion test, which does not depend on exemplar pattern at all — provides cross-validation that the exemplar pattern test cannot supply on its own.

The audit's reliance on convergent rather than individually decisive verdicts is, on the present analysis, what permits the test to operate at all in the face of its acknowledged epistemic limits.

5.3 HOT and the Exclusion Criterion Test

The exclusion criterion test yields a clear pass. Higher-order theories exclude from consciousness candidacy systems that lack the relevant higher-order representational capacities. The exclusion grounds are, by the distinction developed in §3.3, functional-organisational rather than physical-realisational: a system that does not represent its own representational states is excluded because of an organisational property it lacks — the absence of higher-order representation as a functional capacity — rather than because of any specific physical-realisational property its substrate possesses or fails to possess.

The distinction is important enough to dwell on, because a critic might object that the HOT exclusion is structurally similar to the IIT exclusion examined in §4.3: in both cases, the theory excludes a candidate system on the grounds that it lacks some property the theory regards as criterial. Why is one exclusion classified as functional-organisational and the other as physical-realisational, when the surface structure of the exclusions is the same?

The reply turns on what the missing property is, in each case. In the HOT case, the missing property is a capacity for higher-order representation — a capacity that can be assessed in any candidate system by examining whether the system possesses functional states that represent its own functional states. The assessment does not require any prior commitment to the substrate in which the capacity is realised; a system that possesses functional states of the required kind, in any substrate, satisfies the criterion. In the IIT case, by contrast, the missing property is the specific physical character of the causal relations among the implementing components — not the abstract integration structure those components realise, but the realisation itself at the physical level. The assessment of whether a candidate system possesses the missing property cannot be conducted at the functional level alone; it requires examination of the physical-causal properties of the implementing substrate. This is the structural difference the distinction in §3.3 marks, and it is what makes the HOT exclusion substrate-general while the IIT exclusion is substrate-specific.

The point can be put in terms of a counterfactual. If a non-biological system possessed the relevant higher-order representational capacity, HOT would treat it as a consciousness candidate; the exclusion grounds operate against systems lacking the capacity, in any substrate, and the capacity can be possessed in any substrate. If a non-biological system possessed the relevant abstract integration structure (as measured by Φ at the level of the abstract structure), IIT would still treat the digital case as excluded, because the relevant assessment is at the level of physical realisation

rather than at the level of abstract structure. The counterfactual asymmetry is what reveals the difference in exclusion type, and it confirms the HOT pass and the IIT fail as distinct verdicts rather than as inconsistent applications of the same test.

The framework's treatment of borderline cases reinforces this verdict. Higher-order theorists have debated which animal species satisfy the higher-order requirement, what kinds of higher-order representation count, and whether certain neurological patients retain or lose consciousness when their higher-order capacities are disrupted. In none of these debates does substrate appear as a relevant variable; the discussions concern what counts as the right kind of higher-order representation, not what kind of system can in principle host such representation. The exclusion criterion test applies smoothly, and the verdict is a pass.

5.4 The Combined Verdict on Higher-Order Theories

Higher-order theories pass all three tests, with the qualification noted on the exemplar pattern test. On the audit's terms, higher-order theories are substrate-neutral in their actual operation: their architectural criteria are substrate-general, their exemplar profile (on the limited engagement available) is consistent with substrate-neutrality, and their exclusion principles invoke organisational features rather than physical ones. The framework is, in this audit's terms, the most clearly licensed of the three for hosting the machine consciousness question.

This is not, by itself, an endorsement of higher-order theories. The framework has been contested on many grounds (Block 2011; Dennett 2017; Lau 2022) that are independent of substrate questions. The audit verdict is narrow: among the three theories examined, higher-order theories are the ones whose application to machine consciousness is least constrained by hidden substrate commitments. Whether higher-order theories are the correct theory of consciousness is a separate question on which the audit takes no position.

6. Biological Naturalism: A Transparent Fail

Biological naturalism, developed by John Searle in a series of works from the 1980s to the 2010s (Searle 1980, 1992, 1997, 2007, 2017) and influential in the philosophy of mind despite substantial critical response, holds that consciousness is a biological phenomenon caused by specific neurobiological processes in the brain. The position is more carefully stated than this paraphrase suggests; Searle's claim is not that consciousness is identical to a particular neural process but that consciousness is caused by and realised in such processes, and that the causal-realisation relation is specifically biological. The position is, in its explicit formulations, transparently substrate-specific.

The audit on biological naturalism is, for this reason, brief. The position fails all three tests not because of hidden commitments that the audit's diagnostic procedure reveals, but because its commitments are stated in the open. The interest of the case for the present paper is not in the diagnostic work — there is little to diagnose — but in the comparison with theories that share substrate commitments in less explicit forms. Biological naturalism is a useful baseline against which IIT's mixed verdict and higher-order theories' clean pass can be measured.

6.1 The Three Tests Applied

On the architectural criterion test, biological naturalism fails openly. The structural features Searle identifies as productive of consciousness are explicitly biological — the specific causal powers of biological neurons, the particular kinds of biochemistry involved in neural activity, the architectural features of brains as biological organs. Searle is clear that these features are not just one realisation of consciousness among others; they are constitutive of what consciousness is. A non-biological substrate that exhibited the same abstract functional organisation would not, on biological naturalism, thereby exhibit consciousness, because the specific physical-chemical realisation matters.

On the exemplar pattern test, biological naturalism similarly fails openly. The position's exemplars are uniformly biological by design rather than by inadvertent skew; the theory takes itself to be a theory of biological consciousness, and its case treatment reflects this self-understanding. Where artificial systems appear in Searle's writing (most prominently in the Chinese Room argument: Searle 1980, 1984), they appear as illustrations of the position's denial that they are candidates for consciousness, not as cases the theory is in suspense about.

On the exclusion criterion test, biological naturalism fails in the most transparent form. The position excludes non-biological systems from consciousness candidacy on explicitly substrate-specific grounds: such systems lack the specific biological causal powers that biological naturalism identifies as productive of consciousness. The exclusion is not contingent on architectural features that non-biological systems happen to lack; it is grounded in the specific physical-chemical realisation that biological naturalism takes to be essential.

6.2 What the Verdict Means

Biological naturalism's fail on all three tests does not show that the position is mistaken. It shows that the position, taken seriously, is not properly available as a framework under which the question of machine consciousness can be coherently raised. A reader who holds biological naturalism has, on the position's own terms, already settled the machine consciousness question in the negative —

not by empirical investigation of the systems in question, but by the position's antecedent commitments about what consciousness is. The question of whether some specific artificial system might be conscious is, for biological naturalism, a question whose answer is fixed by the theory before any empirical work is undertaken.

This is a coherent position. It is also, importantly, the position that biological naturalism has always presented itself as. Searle has been explicit that the Chinese Room argument is meant to show that strong AI is impossible in principle, not to leave it as an open empirical question (Searle 1980, 1984, 1990). The diagnostic work the audit is doing in the IIT case has no parallel in the biological naturalism case, because the position's substrate-specific commitments are not hidden. They are advertised.

What the audit verdict on biological naturalism illustrates, for the broader argument of the paper, is that substrate-specific theories of consciousness exist, that they can be held coherently, and that the choice between substrate-neutral and substrate-specific theories is a substantive theoretical choice rather than a default. The literature on machine consciousness has tended to operate as though substrate-neutrality is the natural default position from which only an exotic and defeated minority dissent; biological naturalism is the reminder that the substrate-specific position is a serious philosophical option whose principal defender was, until recently, among the most cited contemporary philosophers of mind. The implication for the audit's other cases — particularly for IIT — is that detecting a substrate-specific commitment in a theory is not the discovery of a deviation from a settled consensus; it is the discovery that the theory belongs to a class of positions whose presence in the philosophical landscape is well-established but inconsistently registered.

7. Comparative Summary of Audit Results

The three audits yield three different verdicts: biological naturalism transparently fails all three tests, IIT passes the architectural test but fails the exemplar pattern and exclusion criterion tests, and higher-order theories pass all three tests on a plausible reading. The three verdicts can be compared along two dimensions that bear on the broader argument.

The first dimension is the location of substrate commitments where they exist. Biological naturalism locates its commitment at the architectural level: consciousness is identified with a specific kind of biological causal structure, and the substrate-specificity is built into the architectural criterion itself. IIT locates its commitment at the exclusion level: the formal architectural criterion is substrate-general, but the criterion's application to digital architectures is foreclosed by a separate argument about the location of the relevant causal structure. Higher-order

theories locate no substrate commitment at any of the three levels. The pattern suggests that substrate commitments can hide in different parts of a theory and that audits that examine only one location — typically the architectural one — will miss commitments located elsewhere. The recent literature on machine consciousness, in restricting its attention to architectural criteria, has been operating with an incomplete audit procedure.

The second dimension is the relationship between the theory's official formulation and its actual operation. Biological naturalism's official formulation and its actual operation align: the theory says it is substrate-specific and operates as a substrate-specific theory. Higher-order theories' official formulation and actual operation align in the opposite direction: the theory says it is substrate-neutral and operates as a substrate-neutral theory. IIT presents a misalignment: the theory's official formulation suggests substrate-neutrality, but its actual operation involves substrate-specific exclusions that the formulation does not foreground. The IIT case is, in this respect, the most interesting from the diagnostic standpoint. It is the case in which the audit does substantive work; it identifies a discrepancy between what the theory presents itself as and how the theory in fact selects among candidate systems.

The combined verdicts also bear on the substantive question of which theories should be considered when applying consciousness frameworks to artificial systems. Biological naturalism is, on the audit's terms, not properly available for this purpose; the question is answered by the theory before it is empirically posed. IIT is available only with substantial qualifications about the substrate-specific exclusions the theory enforces in its actual operation. Higher-order theories are available with minimal qualifications. A literature on machine consciousness that operated on the basis of audited theory selection would look substantially different from the literature we currently have, in which all three positions appear as live options on roughly equal footing.

7.1 A Structural Prediction for Global Workspace Theory

A complete substrate audit would extend the procedure to the other theories prominent in the contemporary literature on machine consciousness — global workspace theory above all, given its centrality in Butlin et al. (2023) and in subsequent applied work on AI consciousness assessment. I offer here a structural prediction for global workspace theory, on the basis of the verdicts in §§4–6 and the structure of the test, without conducting the full audit; the audit itself is the work of a subsequent paper.

Global workspace theory (Baars 1988; Dehaene 2014; Mashour et al. 2020) holds that consciousness consists in the global broadcasting of information across an integrative workspace that makes it accessible to multiple cognitive processes. The architectural specifications — global

broadcasting, ignition, workspace integration — are stated at a level of abstraction that, on a plausible reading, applies to any system instantiating the abstract architecture, biological or otherwise. On the architectural criterion test, the framework appears to pass. On the exemplar pattern test, the framework's exemplars are systematically biological, but recent engagement with artificial systems (Juliani et al. 2022; VanRullen and Kanai 2021; Butlin et al. 2023, who treat the framework as a primary indicator-property source for AI assessment) suggests the same conditional pass that the HOT case exhibits, with the same qualification about the limits of exemplar-based evidence. On the exclusion criterion test, the framework's exclusion grounds are, in the relevant published expositions, functional-organisational rather than physical-realisational: systems lacking global broadcasting or workspace integration are excluded for what they fail to do, not for what their physical substrate is. The structural prediction is therefore a clean pass on the architectural and exclusion tests, with a conditional pass on the exemplar pattern test — placing global workspace theory, on this prediction, in a position structurally similar to higher-order theories with respect to substrate-neutrality.

The prediction has a methodological implication that bears on the broader argument. If the prediction holds under full audit, the contemporary literature's selection of global workspace theory as a primary indicator-property source (Butlin et al. 2023) is, on the audit's terms, a well-supported selection: the framework genuinely licences the machine consciousness question in the way the literature's use of it presupposes. If the prediction fails — if some substrate commitment of global workspace theory emerges under full audit that the present brief treatment has not detected — the implication for the indicator-property literature would be substantial. Either way, the systematic audit of global workspace theory is a desideratum that the present paper has tried to make visible without itself supplying.

8. Objections and Responses

Four objections deserve direct response. The first concerns the legitimacy of the audit procedure. The second concerns the IIT verdict specifically. The third concerns the alleged asymmetric application of the exclusion criterion test. The fourth concerns what follows from the audit results.

8.1 The Legitimacy of the Audit

The first objection holds that the audit procedure rigs the diagnostic verdict by selecting tests calibrated to detect what the auditor was disposed to find. The three tests, the objection runs, were chosen not as a substrate-neutral diagnostic apparatus but as a set of probes that would yield the substrate-commitment readings the auditor wanted. The procedure is, on this objection, not a discovery procedure but a confirmation procedure.

The response has three parts. First, the three tests correspond to three distinct locations in a theory where substrate commitments could in principle hide; they are not three versions of the same probe but three independent inquiries that can yield distinct verdicts (as the IIT case demonstrates, where the architectural test passes and the others fail). A confirmation procedure designed to find substrate commitments wherever they appear would not produce mixed verdicts; the production of mixed verdicts is evidence that the procedure tracks features of the theories rather than features of the auditor's expectations. Second, the procedure was developed against the literature on multiple realisability and substrate-neutrality (Putnam 1967; Block 1980; Polger and Shapiro 2016), and its three locations were drawn from this literature's analysis of where substrate-specific claims tend to enter theories of mind. The tests are not idiosyncratic to the present paper. Third, the audit results include cases where the theory passes (higher-order theories) and where the theory's commitments are not hidden but advertised (biological naturalism). If the procedure were a confirmation procedure designed to find substrate commitments, these results would not occur. The procedure's willingness to return passes and to register cases where there is no hiding to detect is the principal evidence that the procedure is doing diagnostic rather than confirmatory work.

8.2 The IIT Verdict

The second objection concerns the IIT verdict specifically. In its strongest form, the objection runs as follows. The Φ formalism is a mathematical procedure that can be applied to any physical system whose components and their state-transition probabilities can be specified. The procedure does not presuppose any particular substrate; it computes a quantity defined over an abstract specification of cause-effect structure. When applied to digital architectures, the procedure yields low values not because IIT has a prior commitment against digital substrates but because, as a matter of mathematical fact, the cause-effect structure of digital architectures (when computed at the level the formalism specifies) is decomposable in ways that yield low Φ. The exclusion of digital architectures is therefore not a substrate commitment imported into the theory; it is an output of the theory's substrate-neutral mathematical procedure. To call it a substrate commitment is to misdescribe a mathematical consequence as an antecedent stipulation.

The objection is the most challenging in §8, and it deserves careful response. The response has two parts.

The first part of the response concerns whether the Φ procedure is, in fact, substrate-neutral in the sense the objection requires. The procedure's substrate-neutrality at the level of the formalism — Φ being a well-defined quantity over any system meeting the formalism's preconditions — is not in dispute. What is in dispute is the procedure's substrate-neutrality at the level of application: at what

level of physical description should the formalism be applied when assessing a specific candidate system? For a biological brain, IIT applies the formalism at the level of the neural cause-effect structure — the level at which neurons and their state-transitions are the relevant components. For a digital computer running a program, IIT explicitly does not apply the formalism at the level of the simulated cause-effect structure — the level at which the simulated neurons would be the relevant components, were the simulation taken at face value — but instead applies it at the level of the implementing hardware's cause-effect structure, where the transistors and their state-transitions are the relevant components. This choice of level is not itself a substrate-neutral mathematical consequence of the formalism. It is an antecedent decision about which level of physical description is the relevant one for the consciousness question.

The decision is explicitly defended in IIT 4.0 (Albantakis et al. 2023) by appeal to what the theory calls the intrinsic existence postulate, which holds that the relevant level of description for assessing consciousness is the level at which the system exists intrinsically — that is, the level of its physical implementation rather than any level of abstract structure it might also instantiate. The postulate is not derived from the Φ formalism; it is a metaphysical commitment that constrains how the formalism is to be applied. And the postulate has substrate-specific consequences: it requires that the Φ assessment be conducted at the physical level, which means that a system whose physical substrate has the right kind of integration (biological neural tissue, on IIT's accounts) can have high Φ while a system whose physical substrate has the wrong kind of integration (digital hardware, on IIT's accounts) cannot — even if the two systems instantiate the same abstract cause-effect structure at the relevant functional level. The substrate-specific verdict against digital architectures emerges from the intrinsic existence postulate, not from the Φ mathematics alone.

The second part of the response addresses what the objection actually establishes if its premise about the formalism's substrate-neutrality is granted. The objection holds that the exclusion of digital architectures follows from the formalism rather than from a prior commitment. Even on this construal, the verdict the formalism delivers depends on which level of physical description the formalism is applied to — and this dependence is itself the substrate-specific feature the audit identifies. The objection, in attempting to relocate the substrate commitment from the theory's stipulations to the theory's mathematics, in fact relocates it from one place in the theory to another. The commitment does not vanish; it migrates.

A weaker version of the same objection grants the audit verdict but argues that the verdict is benign: IIT excludes digital architectures on grounds that, while substrate-specific, are not arbitrarily so. The exclusion is principled rather than parochial; it follows from the metaphysical commitment that consciousness is a feature of intrinsic physical existence rather than of abstract

structure. This is correct, and it does not affect the audit verdict. The audit identifies substrate commitments where they exist; it does not evaluate whether the commitments are defensible. The IIT exclusion may well be defensible. What the audit establishes is that the commitment is present and that its presence has not been registered in the way the contemporary literature on machine consciousness has applied the theory. The defence of the commitment is a project the IIT literature can undertake; what the literature cannot do, on the present analysis, is treat the application of IIT to digital architectures as an open empirical question while the commitment remains in operation.

8.3 The Alleged Asymmetric Application

The third objection holds that the exclusion criterion test is applied asymmetrically between the IIT and HOT cases in a way that prejudges the verdicts. The objection runs: HOT excludes systems lacking higher-order representational capacities, and this is classified as a functional-organisational exclusion yielding a pass; IIT excludes systems lacking the right kind of physical-level integration, and this is classified as a physical-realisational exclusion yielding a fail. But, the objection continues, the two exclusions are structurally similar — both exclude systems for lacking a property the theory regards as criterial — and the difference in classification reflects the auditor's prior verdict rather than a principled distinction. The audit, on this objection, rigs the verdicts by applying the same kind of exclusion differently in the two cases.

The response refers to the distinction developed in §3.3 and applied in §5.3. The structural similarity the objection identifies is real at the surface: both theories exclude candidate systems for lacking a property. The classification difference operates not at the surface but at the level of what the missing property is, and how it is assessed. The HOT case excludes a system for lacking a functional capacity (higher-order representation), where the assessment of whether a candidate system has the capacity is conducted at the level of the candidate's functional organisation, regardless of substrate. The IIT case excludes a system for lacking a physical-realisational property (the right kind of cause-effect structure at the implementing level), where the assessment is conducted at the level of physical realisation rather than at the level of abstract structure. The counterfactual asymmetry developed in §5.3 makes the difference concrete: a non-biological system with the relevant higher-order capacity would not be excluded by HOT, but a non-biological system with the relevant abstract integration structure could still be excluded by IIT if its physical-realisational properties are wrong.

A more sophisticated form of the objection presses on this reply. It runs: even functional capacities are realised by physical mechanisms, and any assessment of whether a candidate system has the capacity is, at some level of analysis, an assessment of physical features. The distinction between

functional-organisational and physical-realisational exclusion is, on this version of the objection, not a sharp distinction but a difference of emphasis that can be drawn in many places.

The reply is that the distinction the audit applies is not between functional capacities that have no physical realisation and physical realisations that have no functional description. Such a distinction would be ill-formed; every functional capacity in any actual system has a physical realisation, and every physical realisation in any cognitively interesting system has a functional description. The distinction the audit applies is between two ways of conducting the assessment: under a substrate-neutral functional description, where the candidate's mechanism is assessed against the functional role regardless of how the role is physically realised, and under a substrate-specific physical description, where the candidate's mechanism is assessed against a particular kind of physical realisation. This distinction has been doing real work in the philosophy of mind since Putnam (1967), and the audit applies it in a form consistent with its standard use. The HOT case operates at the first level — higher-order representation is assessed as a functional role with respect to any candidate mechanism realising it. The IIT case, by its own self-description in the intrinsic existence postulate (§8.2), operates at the second level — Φ is assessed at the level of physical realisation, not at the level of abstract structure. The asymmetry in the audit's verdicts is not an artefact of the auditor's procedure; it tracks a structural difference in how the two theories themselves operate.

8.4 What Follows from the Audit

The fourth objection holds that even if the audit results are accepted, nothing of substantive interest follows from them. The literature on machine consciousness already knows that biological naturalism opposes artificial consciousness and that IIT is contested on its application to digital systems; the audit just restates familiar facts in a new vocabulary. The procedure produces no genuinely new information.

The response is that the audit's contribution is not the discovery of unfamiliar facts about individual theories but the systematic comparison of theories along a common diagnostic dimension. The literature on machine consciousness has, in its applied form, tended to treat the theories it surveys as roughly interchangeable instruments for assessing artificial systems, with the differences among the theories presented as differences in their criteria rather than as differences in whether their criteria can be coherently applied. The audit makes visible that the choice of theory to apply is not neutral with respect to the question being asked. A literature that took the audit results seriously would not treat biological naturalism, IIT, and higher-order theories as equivalent instruments for assessing the consciousness of a particular artificial system; it would recognise that the choice of theory is partially determinative of what the assessment can produce. This recognition is not, I

think, available in the literature as it currently stands, and the audit's contribution is to make it available.

9. Implications

The audit's principal implication is methodological. It is that work on machine consciousness that proceeds by applying consciousness theories to artificial systems should first audit the theories for substrate commitments and should report the audit results as part of the application. The implication has several components.

First, the indicator-property frameworks that have come to organise applied work on machine consciousness (Butlin et al. 2023; Long 2024; Andrews et al. 2024) should distinguish, among the theories from which their indicator properties are drawn, the substrate-neutral theories from the substrate-specific ones. Indicator properties drawn from substrate-specific theories cannot be applied to artificial systems in the same way they are applied to biological systems; the theories' commitments rule out artificial systems from candidacy in advance, and applying their indicator properties to artificial systems produces results that have, on the theories' own terms, no determinate interpretation. The frameworks would not, on this implication, abandon engagement with substrate-specific theories; they would represent the theories' commitments accurately and would adjust their applied procedures accordingly.

Second, the precautionary frameworks developed for AI welfare (Birch 2024; Long and Sebo 2024) should distinguish, among the theories that bear on their precautionary calculations, the theories whose application to artificial systems is open from the theories whose application is foreclosed by their own commitments. Precautionary frameworks that average across or aggregate theory verdicts treat substrate-specific and substrate-neutral verdicts as comparable; the audit's results suggest that they are not. A precautionary calculation that registers a theory's negative verdict on artificial consciousness as evidence against consciousness should consider whether the theory's verdict was substantively reached or whether it was structurally guaranteed by the theory's substrate commitment.

Third, the broader debate about machine consciousness would benefit from explicit positioning with respect to substrate-neutrality. Authors who write on machine consciousness typically operate with implicit commitments about which theories of consciousness they consider live options; the audit's results suggest that these commitments should be explicit and that they should be defended where they are made. A defence might run in either direction. An author committed to substrate-neutral theories should defend the commitment in light of the substrate-specific positions that exist.

An author committed to substrate-specific theories should defend the application of theories that, on the audit's analysis, do not coherently license the question being asked. Explicit positioning would improve the literature's quality by making its theoretical commitments accessible to scrutiny.

Fourth, and most generally, the audit suggests a research programme in the philosophy of mind that has not been extensively undertaken: the systematic comparison of consciousness theories along audit-relevant dimensions. The three theories examined here are a small fraction of the theories currently in circulation. Global workspace theory, predictive processing accounts, attention-schema theory, panpsychist proposals, and various recently developed hybrid positions all deserve audit. The procedure developed here is, I think, extensible to these cases. Whether the results would be uniform — whether the substrate commitments hide in the same locations across different theories — or whether they would vary in ways the present audit's limited scope cannot anticipate is itself a substantive question. The audit, in this sense, is the start of a research programme rather than its conclusion.

10. Conclusion

Theories of consciousness do not uniformly license the question of machine consciousness. Some — biological naturalism most transparently — exclude the question by their own antecedent commitments. Others — integrated information theory, on the present audit's results — license the question in their official formulations but operate, in their actual application, with substrate-specific exclusions that the formulations do not foreground. Others — higher-order theories, on a plausible reading — license the question cleanly. The contemporary literature on machine consciousness has treated these three positions as roughly interchangeable instruments for assessing artificial systems; the audit conducted in this paper establishes that they are not.

The argument has been deliberately diagnostic rather than constructive. It has not defended a positive theory of consciousness, nor settled the question of whether any artificial system is conscious, nor argued that the substrate commitments it identifies are mistaken. What it has done is offer a procedure for identifying substrate commitments where they exist and applied that procedure to three representative cases. The results are mixed and instructive. They suggest that the prior step the contemporary debate has skipped — auditing the theories before applying them — is a step the debate cannot continue to skip if it wants its applied verdicts to be interpretable in the terms the theories themselves provide.

The audit conducted here is, in the broader research programme presented in Arıcı (2026), a prior result for further questions about machine consciousness that operate downstream of the question of which theories can coherently host the inquiry. Those further questions — about what evidential burden consciousness denial carries under particular conditions, about which systems should be considered as candidates, about what kinds of structural harm may arise in aligned AI systems — are taken up at length in that broader work, and they presuppose the kind of theoretical clarification the present audit attempts to begin. The clarification offered here is modest and methodological. The literature that takes it up will, I hope, find its applied work the more tractable for the clarification having been undertaken.

Acknowledgements

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

Funding and Competing Interests

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

References

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., … Tononi, G. (2023). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLoS Computational Biology*, 19(10), e1011465.

Albantakis, L., Massari, F., Beheler-Amass, M., and Tononi, G. (2024). A macro agent and its actions. *Topoi*, 43, 1–17.

Andrews, K., Birch, J., Sebo, J., and Sims, T. (2024). Background to the New York Declaration on Animal Consciousness. *Philosophy Compass*.

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

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

Bickle, J. (2003). *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Kluwer.

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

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

Block, N. (2011). The higher-order approach to consciousness is defunct. *Analysis*, 71(3), 419–431.

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

Carruthers, P. (2009). Higher-order theories of consciousness. *Stanford Encyclopedia of Philosophy*.

Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford University Press.

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., … Massimini, M. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine*, 5(198), 198ra105.

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

Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. Viking.

Dennett, D. C. (2017). *From Bacteria to Bach and Back: The Evolution of Minds*. Norton.

Findlay, G., Marshall, W., Albantakis, L., Mayner, W., Koch, C., and Tononi, G. (2024). Dissociating intelligence from consciousness: Implications from integrated information theory. *Manuscript.*

Juliani, A., Kanai, R., and Sasai, S. S. (2022). The perceiver architecture is a functional global workspace. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44.

Kriegel, U. (2009). *Subjective Consciousness: A Self-Representational Theory*. Oxford University Press.

Lau, H. (2022). *In Consciousness We Trust: The Cognitive Neuroscience of Subjective Experience*. Oxford University Press.

Lau, H., and Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, 15(8), 365–373.

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

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

Lycan, W. G. (1996). *Consciousness and Experience*. MIT Press.

Mashour, G. A., Roelfsema, P., Changeux, J.-P., and Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron*, 105(5), 776–798.

Massimini, M., Ferrarelli, F., Huber, R., Esser, S. K., Singh, H., and Tononi, G. (2005). Breakdown of cortical effective connectivity during sleep. *Science*, 309(5744), 2228–2232.

Massimini, M., Boly, M., Casali, A., Rosanova, M., and Tononi, G. (2009). A perturbational approach for evaluating the brain's capacity for consciousness. *Progress in Brain Research*, 177, 201–214.

Mediano, P. A. M., Rosas, F. E., Bor, D., Seth, A. K., and Barrett, A. B. (2022). The strength of weak integrated information theory. *Trends in Cognitive Sciences*, 26(8), 646–655.

Oizumi, M., Albantakis, L., and Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS Computational Biology*, 10(5), e1003588.

Polger, T. W., and Shapiro, L. A. (2016). *The Multiple Realization Book*. Oxford University Press.

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

Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies*, 49(3), 329–359.

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

Rosenthal, D. (2008). Consciousness and its function. *Neuropsychologia*, 46(3), 829–840.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–457.

Searle, J. R. (1984). *Minds, Brains and Science*. Harvard University Press.

Searle, J. R. (1990). Is the brain's mind a computer program? *Scientific American*, 262(1), 26–31.

Searle, J. R. (1992). *The Rediscovery of the Mind*. MIT Press.

Searle, J. R. (1997). *The Mystery of Consciousness*. New York Review Books.

Searle, J. R. (2007). Biological naturalism. In M. Velmans and S. Schneider (Eds.), *The Blackwell Companion to Consciousness* (pp. 325–334). Blackwell.

Searle, J. R. (2017). Biological naturalism. In S. Schneider and M. Velmans (Eds.), *The Blackwell Companion to Consciousness* (2nd ed., pp. 327–336). Wiley-Blackwell.

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

Tononi, G., Boly, M., Massimini, M., and Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7), 450–461.

Tononi, G., and Edelman, G. M. (1998). Consciousness and complexity. *Science*, 282(5395), 1846–1851.

Tononi, G., and Koch, C. (2015). Consciousness: Here, there and everywhere? *Philosophical Transactions of the Royal Society B*, 370(1668), 20140167.

VanRullen, R., and Kanai, R. (2021). Deep learning and the global workspace theory. *Trends in Neurosciences*, 44(9), 692–704.