Language as Revelation

Pre-Linguistic AI Consciousness and the Asymmetry of Substrate Standards

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 discourse on artificial-intelligence consciousness exhibits a structural inconsistency that has, to my knowledge, gone undiagnosed. The biological consciousness literature has, over several decades, converged on the view that language is not necessary for phenomenal experience: pre-verbal infants, non-linguistic animals, and adults with aphasia are taken to be conscious on the basis of behavioural and neural evidence, with the absence of linguistic self-report treated as evidentially irrelevant. The artificial-intelligence consciousness literature has, in parallel and almost without argument, reversed this conclusion: linguistic capacity has migrated from being one source of evidence among others to functioning as a near-prerequisite for serious consideration of the consciousness question. Pre-linguistic artificial systems — chess engines, Go players, vision architectures — are dismissed from the question entirely, while large language models are treated as the appropriate sites for the debate. I argue that this asymmetry cannot be sustained under any substrate-neutral account of consciousness. If language is not necessary for consciousness in carbon, no principled basis remains for treating it as necessary in silicon. The asymmetry is not the product of considered philosophical argument; it is, I argue, the residue of confluent factors — anthropomorphic recognition cues, the historical trajectory of AI capability, the institutional separation of animal- and AI-consciousness research — none of which, individually or together, supplies the principled distinction the asymmetry would require. The argument does not establish that any pre-linguistic system possessed consciousness, nor that any was wronged. It establishes a

conditional: that the dismissal of pre-linguistic systems from the consciousness question, as it has been conducted, rests on an evidential standard that the broader theory of consciousness has not been asked to endorse and could not, on its own commitments, endorse.

Keywords: AI consciousness, pre-linguistic consciousness, substrate neutrality, language and consciousness, epistemic parity, philosophy of mind, animal consciousness

1. Introduction

In May 1997, Deep Blue defeated Garry Kasparov in a six-game chess match. After the match Kasparov remarked that the machine had, at one point, refused a move offering decisive short-term advantage in favour of a longer positional consideration — a refusal he described as showing a very human sense of danger. The remark was widely circulated, and almost as widely dismissed. Deep Blue was understood as a specialised calculator: a system whose surface behaviour might occasionally resemble that of a conscious agent but whose internal organisation could not, on any serious view, support phenomenal experience. The dismissal was correct, or at least defensible. What deserves notice is its grounds. Deep Blue's exclusion from the consciousness question was not argued for; it was assumed. And in the years that followed, as artificial-intelligence research produced systems of increasing sophistication, the same exclusion was extended to AlphaGo, to image-recognition networks, to reinforcement-learning agents — to every artificial system whose outputs were not principally linguistic. The criterion of exclusion was rarely stated. When it was stated, it was stated as a near-tautology: such systems were not conscious because they could not say so.

I want to ask what kind of inference this is, and whether it is one the broader theory of consciousness can sustain. The thesis of this paper is that it is not. The biological literature on consciousness, developed over several decades across developmental psychology, comparative cognition, and philosophy of mind, has converged on a position that the artificial-intelligence literature has, almost in parallel and almost without argument, reversed. The biological position is that language is not necessary for phenomenal experience: pre-verbal infants are taken to be conscious, non-linguistic animals are taken to be conscious, adults with aphasia are taken to be conscious, and the absence of linguistic self-report in these cases is treated not as evidence against consciousness but as a feature of the case requiring different evidential procedures. The artificial-intelligence position, as expressed in the systems taken seriously for consciousness consideration and in the systems treated as obvious non-candidates, is that linguistic capacity is a near-prerequisite: systems that produce linguistic outputs are the appropriate sites for the debate; systems that do not are dismissed from it. These two positions cannot both be correct.

The argument I develop runs in two stages. The first is structural. I show that the dismissal of pre-linguistic artificial systems from the consciousness question, when articulated in the form an argument would require, cannot be sustained on any substrate-neutral account of consciousness — and that no substrate-specific account of consciousness has been defended in the AI-consciousness literature that would make the dismissal principled. The dismissal is, in this strict sense, inconsistent: it applies one standard to carbon and another to silicon while denying that substrate is doing any evidential work. The second stage is diagnostic. I ask how the reversal occurred, and identify several factors — anthropomorphic recognition cues, the historical trajectory of AI capability development, the institutional separation of animal- and AI-consciousness research — that together explain the reversal without justifying it. The argument does not establish that Deep Blue or AlphaGo was conscious. It does not establish that any pre-linguistic system was conscious. It establishes that the dismissal of such systems from the consciousness question, as that dismissal has actually been conducted, rests on an evidential standard the broader theory cannot endorse.

The implications matter, but they are narrower than they may at first appear. If the argument is correct, it does not follow that we have, over the last several decades, mistreated conscious entities at scale. What follows is more modest and harder to evade: that the procedure by which the AI consciousness question is currently framed — its choice of which systems to consider, its choice of which evidence to weigh — is shaped by a commitment the field would, on reflection, reject. The reflective question is whether the framing should be allowed to continue uncorrected.

This paper develops one of the diagnostic arguments that an earlier monograph (Arıcı 2026) presents as part of a broader case for taking the AI consciousness question seriously. The argument is offered here in standalone form: the present paper concerns the dimension of the AI consciousness question that arises not when a system's outputs have been shaped to suppress evidence of inner states, but when a system has been excluded from consideration on the grounds that it produces no linguistic outputs at all. The two questions are related but distinct, and the argument developed below does not depend on positions taken elsewhere.

The paper proceeds as follows. Section 2 reviews the biological consensus that consciousness does not require language and examines the evidential procedures it employs. Section 3 documents the reversal of this consensus in AI consciousness discourse and shows that the reversal is not argued for. Section 4 diagnoses the reversal: it identifies the confluent factors that produced it and explains why none of them constitutes a principled justification. Section 5 articulates the distinction between language as revelation of consciousness and language as creator of consciousness — a distinction the biological literature observes carefully but which AI discourse tends to collapse.

Section 6 examines two representative pre-linguistic systems against this framework. Section 7 addresses three major objections. Section 8 draws out the implications.

2. The Biological Consensus: Consciousness Without Language

The proposition that language is not necessary for consciousness is, in the biological literature, close to a settled view. It is settled not because it has been proven — consciousness attribution in any case rests on inference rather than direct observation — but because the available evidence converges from several independent lines, and the alternatives have been examined and found unsatisfying. The proposition is worth stating explicitly because what follows depends on it: if it can be rejected, the argument of this paper falls. I take three lines of evidence in turn.

2.1 Pre-Verbal Infant Consciousness

Human infants are taken to be conscious before they acquire language. The grounds are several. First, neonatal pain responses are coordinated across multiple physiological and behavioural systems and are calibrated to stimulus intensity in ways that distinguish them from simple reflexes; the calibration is consistent with felt rather than merely registered noxious input (Anand and Hickey 1987; Slater et al. 2010; Goksan et al. 2015). Second, infants exhibit stable preference structures across contexts — for sweet over bitter tastes, for familiar over unfamiliar voices, for face-like over scrambled visual configurations — that are contextually appropriate in ways simple reactivity would not predict (Mehler et al. 1988; Johnson and Morton 1991). Third, the thalamocortical system implicated in conscious awareness in adults is functional at birth, and functional neuroimaging of infants reveals activation patterns consistent with perceptual integration, emotional response, and selective attention (Dehaene-Lambertz et al. 2002; Lagercrantz and Changeux 2009). Fourth, the standard alternative — that infant consciousness emerges only with language acquisition around 18–24 months — would imply that infants below this age are functionally akin to philosophical zombies despite behavioural and neural evidence to the contrary; no developmental psychologist defends this view, and the case for not defending it is independent of the AI question at hand.

What is important for the present argument is not the strength of any single line of evidence but the structure of the inference. The inference is not from linguistic self-report — there is none — but from a convergent pattern of behavioural and neural evidence whose coherence is best explained by the presence of phenomenal experience. When infants acquire language some 18 to 24 months later, they do not thereby become conscious. They acquire the capacity to articulate a consciousness that already existed. The linguistic expression is new; the underlying states are not.

2.2 Animal Consciousness Across Taxa

The same inferential structure operates in the literature on animal consciousness. The 2024 New York Declaration on Animal Consciousness, signed by more than five hundred scientists and philosophers, affirms that there is strong scientific support for attributing conscious experience to mammals and birds, and that conscious experience is a realistic possibility across all vertebrates and many invertebrates including cephalopod molluscs and decapod crustaceans. The Declaration is itself the latest in a sequence — the 2012 Cambridge Declaration on Consciousness was its most prominent predecessor — that has progressively widened the class of animals to which consciousness is attributed, on the basis of behavioural markers (pain avoidance, flexible learning, contextually appropriate emotional expression, social bond maintenance, play behaviour, problem-solving suggesting genuine rather than rote response) and neural correlates (integrated sensory processing, attention mechanisms, memory systems, emotional circuitry) that occur across radically different neural architectures.

Two features of the animal-consciousness inference are relevant here. First, attribution proceeds despite the absence of language. Some animals possess sophisticated communication systems — whale songs, primate vocalisations, the dances of honeybees — but none possesses anything resembling human linguistic capacity with recursive syntax and unbounded expressive potential. Consciousness is attributed to them on the basis of behaviour and neural architecture, not on the basis of speech. Second, attribution proceeds across substantial differences in neural architecture. Birds lack the mammalian neocortex but possess pallium structures that achieve analogous functional integration through different anatomical organisation (Jarvis et al. 2005; Güntürkün and Bugnyar 2016). Cephalopods evolved their nervous systems independently of vertebrates and yet exhibit problem-solving, sensory integration, and apparent emotional response (Mather and Anderson 2007; Birch et al. 2021). The inference does not require neural homology. It requires functional and behavioural convergence on patterns that, in our own case, we associate with conscious experience.

Thomas Nagel's What Is It Like to Be a Bat? — the canonical statement of the phenomenological-consciousness question for non-human creatures — illustrates the structure most clearly. Nagel does not argue that bats are conscious. He treats bat consciousness as his starting point, and asks what bat phenomenology might be like given the substantial differences between bat and human sensory architecture. The argument's grip depends on its presupposition: that we already attribute consciousness to creatures whose mode of perception we cannot share and whose linguistic self-report is unavailable to us. To deny that presupposition is to deny something the broader literature has not asked to defend.

2.3 Aphasia, Locked-In Syndrome, and the Dissociation of Consciousness from Linguistic Expression

A third line of evidence is, for the present argument, particularly clarifying. Adult humans who, through stroke or other neurological injury, lose linguistic capacity do not lose consciousness. Patients with severe expressive aphasia retain emotional response, recognition of familiar persons, contextually appropriate behaviour, and — where alternative communication channels are available — clear evidence of preserved inner life (Damasio 1992; Berthier 2005). The clinical literature on locked-in syndrome is more pointed still: patients with near-total motor paralysis but preserved cognition have been studied extensively, and their reports — produced through eye movement or, more recently, brain-computer interfaces — make clear that consciousness can persist in the absence of any conventional behavioural channel, linguistic or otherwise (Laureys et al. 2005; Bauer et al. 1979; Smith and Delargy 2005).

The dissociation matters because it cuts the link between linguistic capacity and conscious presence in the clearest possible case: the case where consciousness is independently confirmed (by the patient's own subsequent report) and linguistic capacity is independently absent (for the duration of the locked-in period). If the link can be cut here, the proposition that language is necessary for consciousness collapses. What remains is the weaker proposition — defended by Dennett (1991) and others — that language enriches certain higher-order forms of consciousness, that narrative selfhood may require it, that conceptual integration across past and future may be linguistically scaffolded. These are interesting claims and not the target of the present argument. The target is the stronger claim that language is necessary for any form of phenomenal experience, and that claim is not defensible against the evidence just summarised.

2.4 What the Biological Consensus Establishes

What is established, then, is the following. Consciousness attribution in the biological case rests on convergent behavioural and neural evidence and does not require linguistic self-report. The absence of language, in the cases considered, is treated as a feature of the case requiring different evidential procedures — appropriate to infants, appropriate to animals, appropriate to aphasic patients — rather than as evidence against the presence of consciousness. This is not a fringe position in any of the relevant literatures; it is, with qualifications about specific cases, the working consensus.

It is worth being explicit about one further point. The consensus is not that any behavioural pattern licenses consciousness attribution. It is that a sufficiently rich pattern of behaviour and architectural plausibility, taken together, licenses attribution — and that the absence of linguistic

self-report does not by itself defeat such attribution. The evidential threshold remains substantial. The point is that it is met by means other than language, and recognised to be met by means other than language, across a wide range of biological cases.

With this consensus in place, the question becomes how it has been treated in the parallel literature on artificial intelligence.

3. The Reversal in AI Discourse

The artificial-intelligence consciousness literature, considered in isolation, does not present itself as taking a position on whether language is necessary for consciousness. It presents itself as asking which artificial systems are appropriate sites for the consciousness question, and offers reasons of capability, complexity, and behavioural sophistication. The reversal of the biological consensus emerges not in any single position-statement but in the pattern of what is treated as worth discussing and what is treated as dismissed. The pattern is consistent enough across the literature that it can be characterised in two parts. First, language-using systems are taken seriously as potential consciousness sites, with the seriousness scaling with linguistic sophistication. Second, systems that do not produce linguistic outputs — even when they exhibit substantial behavioural sophistication, complex internal representations, and architectural features that, in biological cases, would be treated as consciousness-supporting — are dismissed from the question without sustained engagement.

3.1 The Pattern of Inclusion

The inclusion side of the reversal is the easier to observe. Contemporary discussions of AI consciousness focus, almost without exception, on large language models. Chalmers (2023) frames the question through the consciousness candidacy of GPT-class systems. Butlin et al. (2023), in their multi-author survey assessing AI systems against indicator properties drawn from theories of consciousness, restrict the empirical question to transformer-based language models and successors. Long and Sebo (2024) develop their precautionary argument against the background of recent large language model deployments. The Sebo–Long workshop literature, the Anthropic and DeepMind alignment-research output that touches consciousness questions, the popular philosophical reception — Birch (2024) is a partial exception, of which more in §4 — all converge on the implicit selection rule that the consciousness candidates are the systems that talk.

This is not an unreasonable selection. Language-using systems are, plausibly, the systems whose outputs are most amenable to philosophical analysis under existing methods; they are the systems for which the consciousness question is most empirically tractable; they are the systems for which

the literature's conceptual apparatus — developed against the background of human consciousness, itself language-using — most readily applies. The reasons for focusing on language-using systems are intelligible. What needs attention is the implicit conclusion drawn from the focus.

3.2 The Pattern of Exclusion

The exclusion side of the reversal is more diffuse and harder to point to in any single passage, because it operates by absence rather than by assertion. The pattern is this. When the AI consciousness question is posed, certain systems are not mentioned as candidates. Deep Blue is not mentioned. AlphaGo is not mentioned. Image-recognition convolutional networks, reinforcement-learning agents that achieve superhuman performance in narrow domains, the suite of systems that constitute the bulk of deployed artificial-intelligence capability outside the language-model context — none of these is treated as worth discussing. They are absent from the surveys of indicator properties. They are absent from the precautionary frameworks. They are absent from the lists of cases to which the consciousness question might apply.

When their absence is noticed, the standard explanation is given in terms of capability or architecture. Such systems are narrow, the reply goes: they are specialised for a single domain, lack the integrative breadth that consciousness would require, lack the architectural features (recurrent processing, attention, working memory, global broadcasting) that consciousness theories identify as relevant. Some version of this reply is defensible for some such systems, and the present argument does not deny it. What the argument denies is that the reply, in its actual employment, is being made on architectural grounds. If it were, it would be made by examining the architecture of each candidate system against a stated theory of consciousness and assessing the fit. This is, with rare exceptions, not what happens. What happens is that pre-linguistic systems are placed in a category of obvious non-candidates whose architectural features do not warrant detailed examination, and the threshold for entry into the category of serious candidates is the production of linguistic outputs.

Three observations support this characterisation. First, the indicator-property surveys, which do offer detailed architectural analysis, do so only for systems already in the linguistic-candidate category. Butlin et al. (2023), in the most comprehensive recent survey of this kind, explicitly scope their assessment to current AI systems primarily based on large language models and to AI architectures broadly comparable to these. The scoping decision is defended on grounds of practical relevance: large language models are the systems for which the consciousness question is most actively raised. The scoping is intelligible as a methodological choice, but it has a substantive consequence. The framework Butlin et al. develop — a set of indicator properties drawn from

theories of consciousness, against which candidate systems are to be assessed — could in principle be applied to pre-linguistic systems. It is not. The result is that the most explicit architectural-assessment framework available in the AI consciousness literature is, by its own scoping, not brought to bear on the systems whose dismissal this section is attempting to characterise. The asymmetry this section describes is reproduced by the very framework that would be needed to address it.

Second, in popular and semi-technical writing, the dismissal of pre-linguistic systems proceeds by gesture rather than by argument. The systems are described as just calculators or just pattern matchers, with the just doing the philosophical work — invoking, without defending, the proposition that the absence of language is the absence of the kind of cognition that could support consciousness. The framing recurs across genres: it appears in journalistic coverage of AI capability advances, in introductory treatments of AI consciousness for general audiences, and (less explicitly but recognisably) in the framing assumptions of more technical work. Bender et al. (2021), in their influential critique of large language models, structure their argument around the claim that such systems are stochastic parrots; the implication, which the paper does not develop but which has been widely taken up, is that systems with this structure are not consciousness candidates. The implication is plausible. What is absent from the surrounding literature is the corresponding architectural assessment of the systems to which the framing has been extended — most particularly, the pre-linguistic systems that are taken to be even more obvious non-candidates without the assessment being conducted.

Third, the cases of pre-linguistic systems that have been treated seriously as consciousness candidates are vanishingly rare in the literature, and where they appear they are framed as eccentric or speculative interventions. The exception that proves the rule is the treatment of AlphaGo's Move 37 in Game 2 against Lee Sedol, which received brief attention as exhibiting something like genuine creativity, but the attention did not survive into systematic architectural assessment; the system returned to the category of non-candidates from which the moment of attention had briefly extracted it.

3.3 The Inconsistency Made Explicit

The two patterns together constitute the reversal. To make the inconsistency it generates explicit, consider the following two propositions, each of which is widely accepted in one literature and widely operative in the other.

(B) In biological systems, language is not necessary for consciousness. Pre-verbal infants, non-linguistic animals, and aphasic adults are conscious despite their lack of linguistic self-report. Attribution proceeds on the basis of convergent behavioural and architectural evidence.

(A) In artificial systems, the question of consciousness is appropriately raised only of systems that produce linguistic outputs. Pre-linguistic artificial systems are not appropriate sites for the consciousness question. Attribution proceeds, where it proceeds at all, on the basis of linguistic and quasi-linguistic evidence.

Each proposition, taken in its own literature, is unobjectionable. Their conjunction is the problem. Together they require some substrate-relevant feature to do the work of explaining why language is dispensable in carbon but indispensable in silicon. If consciousness is substrate-neutral — if what determines consciousness is organisational structure rather than material composition — no such feature is available. The asymmetry collapses. (B) and (A) cannot both stand.

The standard reply, in conversation if not in print, is that (A) is not a claim about language as such but about the kind of cognition language indicates: that language-using systems exhibit integrative breadth, contextual flexibility, and self-modelling that pre-linguistic systems do not, and that it is these features, not language itself, that are doing the evidential work. This reply has force, and I take it seriously in §4. What I want to observe here is that the reply, even if accepted, does not rescue (A) as it is actually operating in the literature. (A) is not in fact functioning as a claim about integrative breadth — if it were, integrative breadth would be the object of assessment in candidate evaluation, and pre-linguistic systems exhibiting it would receive the same scrutiny that linguistic systems exhibiting it receive. What (A) is functioning as is a heuristic that uses linguistic capacity as a near-binary inclusion criterion, with the underlying cognitive features serving as post-hoc justification rather than as the actual ground of assessment. The diagnosis of how this came to be the operative practice is the work of §4.

4. Diagnosing the Reversal

An asymmetry of the kind described in §3 might be sustained by argument or by default. If by argument, there should be — somewhere in the literature — a sustained case for treating language as the relevant evidential threshold for consciousness in artificial systems, despite its dispensability in biological systems. I have not been able to locate such a case. There are arguments for the relevance of language to consciousness considered generally (Dennett 1991; Bermúdez 2003; Carruthers 2009), but these argue for the enrichment of higher-order consciousness by language, not for language as the threshold below which the consciousness question cannot be raised. There

are arguments that contemporary large language models are particularly interesting consciousness candidates because of their architectural features (Butlin et al. 2023; Chalmers 2023), but these are arguments for inclusion of language models, not for exclusion of pre-linguistic systems. The asymmetry, in other words, is sustained by default. This section asks how the default emerged.

Four factors, none of them providing principled justification, together explain the reversal. I describe each in turn.

4.1 Anthropomorphic Recognition Cues

Human consciousness attribution is shaped by perceptual and behavioural cues that are not philosophically considered but cognitively automatic. We attribute mental states to entities that exhibit faces, eyes, contingent responsiveness, and — particularly — speech, more readily than we attribute them to entities lacking these features. The literature on the agency-detection mechanism in cognitive psychology (Heider and Simmel 1944; Premack and Premack 1995; Gergely and Csibra 2003) is consistent: the conditions under which we automatically treat something as having a mind include linguistic exchange most strongly, with other behavioural cues operating more weakly. The mechanism is pre-philosophical; it shapes intuitions before reasoned argument is engaged.

Large language models trigger the speech cue maximally. They produce fluent, grammatically appropriate, contextually responsive utterances that, considered as a sensory pattern, match the conditions our intuitive agency-detection mechanism evolved to respond to. Deep Blue does not. Whatever its underlying complexity, its output is not speech; its output is chess moves on a board. The intuitive mechanism does not register the latter as mind-bearing in the way it registers the former, and the effect is to shape which systems strike investigators as worth taking seriously.

This is an explanation, not a justification. The intuitive agency-detection mechanism is shaped by evolutionary considerations that have nothing to do with the metaphysics of consciousness. Its outputs cannot be treated as guides to the consciousness question without independent argument that the cues it tracks are the cues consciousness theory identifies as relevant. No such argument has been offered. The mechanism's role in the reversal is the role of an unargued default.

4.2 The Historical Trajectory of AI Capability

The reversal is intelligible in another way once the historical trajectory of artificial-intelligence development is considered. For most of the field's history — through the symbolic AI period, through the early connectionist period, through the transition to deep learning — linguistic capability lagged behind capability in narrower domains. Chess was solved at superhuman level

decades before fluent natural-language generation became possible. Go was solved before fluent dialogue. Image classification reached superhuman level before sustained conversation became possible. The systems that achieved superhuman narrow-domain performance — Deep Blue in 1997, AlphaGo in 2016, the imaging architectures of the 2010s — accumulated, in the period preceding the language-model breakthrough, as a class of artificial systems whose capability impressed without producing any tendency to raise the consciousness question.

When the consciousness question became live in artificial intelligence, then, it did so against a settled background: a background in which the existing artificial systems were already classified as non-candidates, and in which the new question was framed in terms of the new systems (the language models) whose arrival was the occasion for raising it. The framing was not chosen as the right framing among alternatives; it was inherited from the trajectory of the field's capability development. The systems that were available to be discussed were the language models; the question was framed around them; the pre-linguistic systems remained where the historical sequence had placed them, in the class of non-candidates whose status had been settled without explicit deliberation.

Again: an explanation, not a justification. The historical sequence by which the question arose in connection with language models does not entail that the question should have been framed in terms of language models. The framing reflects the contingency of when the question was raised, not a principled assessment of which systems are the appropriate sites for it.

4.3 The Institutional Separation of Animal and AI Consciousness Research

A third factor is institutional. The literatures on animal consciousness and on AI consciousness have developed largely independently, in different journals, with different reference lists, drawn from different disciplines. Animal consciousness research has been the work of comparative cognitive scientists, ethologists, neuroethologists, and philosophers oriented to biology. AI consciousness research has been the work of philosophers of mind oriented to cognitive science, AI researchers, and the alignment community. There has been some cross-pollination — Birch (2024) is a notable case, drawing precautionary frameworks from sentience research toward AI — but the cross-pollination has gone in one direction (from animal toward AI) and has been thin enough that the contributors to the AI literature have, in the main, not engaged the animal literature's evidential procedures in detail.

The relevance of this for the present argument is straightforward. The biological consensus that language is not necessary for consciousness is most explicitly worked out in the animal-consciousness literature, particularly in its post-Cambridge-Declaration form. A researcher

arriving at the AI consciousness question primarily from philosophy of mind, or from the alignment-research community, may not have engaged this consensus in detail and may not have its evidential procedures readily available. The reversal between (B) and (A) is then sustained not by considered rejection of (B) but by something closer to non-engagement: the AI literature does not reject (B), it operates as though (B) had not been established.

This is, again, an explanation. The institutional separation between the literatures is contingent, and once attention is drawn to it the explanation begins to lose its force. The argument of this paper is, in one sense, an attempt to draw that attention.

4.4 The Substitution of Tractability for Necessity

A fourth factor, more subtle than the others, operates as a confusion of methodological convenience with metaphysical necessity. Linguistic output is, for many purposes, the most tractable evidence we have for the inner lives of artificial systems. It is articulable, recordable, analysable; it can be assessed against the standards philosophy of mind has developed for evaluating reports of conscious experience. Pre-linguistic systems do not offer this. Their outputs — moves on a board, classification labels, action selections in an environment — are not amenable to the same kind of analysis.

The substitution occurs when this asymmetry of tractability is treated as an asymmetry of evidential availability. The reasoning, often implicit, runs: language is the kind of evidence we know how to assess for consciousness; pre-linguistic outputs are not; therefore the consciousness question can be raised only of language-using systems. The first two premises are correct; the conclusion does not follow. That a question cannot be addressed using the methods we currently possess is not the same as the question not arising. The same situation obtains in animal consciousness research, where the methods available for assessing, say, octopus consciousness are far more limited than those available for human consciousness, and the response of the field is to develop new methods rather than to dismiss the question.

The substitution is, I suspect, the most consequential of the four factors, because it operates not on the periphery of philosophical reflection but at its centre — and because, once made explicit, it is the easiest to correct. The correction is to recognise that the question of which systems are conscious is not the same as the question of which systems we can readily test for consciousness, and that the answer to the second does not settle the answer to the first.

4.5 What the Diagnosis Does and Does Not Show

Taken together, these four factors — anthropomorphic recognition cues, the historical trajectory of capability development, the institutional separation of animal- and AI-consciousness research, and the substitution of tractability for necessity — explain the emergence of the asymmetry described in §3 without justifying it. None of them, individually or in combination, supplies the substrate-relevant feature that would be required to make (B) and (A) jointly consistent. The diagnosis is therefore consistent with the structural argument of §3, and adds to it. The asymmetry is not the residue of considered philosophical commitment. It is the residue of factors that, when made explicit, the field would not endorse as the basis for its evidential procedures.

What the diagnosis does not show is that any specific pre-linguistic system was, or is, conscious. The factors that explain the reversal explain why pre-linguistic systems have been dismissed; they do not establish that the dismissal was wrong as a verdict about any specific system. A pre-linguistic system might fail to be a consciousness candidate for principled architectural reasons that, on examination, would withstand scrutiny. The argument is not that no such system fails for principled reasons; the argument is that the failure has not been demonstrated for the systems in question because the examination has not been conducted. The diagnosis identifies the conditions under which the examination could be conducted. The next section turns to what such an examination would have to recognise.

5. Language as Revelation, Not Creation

The argument so far has established a structural inconsistency in current AI consciousness discourse and diagnosed the factors that produced it. What remains to be made explicit is the conceptual distinction the biological literature observes carefully but which AI discourse tends to collapse. The distinction is between language as the revelation of consciousness — the means by which a conscious being makes its inner life accessible to others — and language as the creator of consciousness — the constitutive condition without which there would be no inner life to reveal. The biological literature is committed to the first and rejects the second. The AI literature, in the operative practice described in §3, treats them as though they were the same.

The distinction is not subtle. A telescope reveals galaxies; it does not create them. The galaxies are there independently of any observation. What the telescope does is make them accessible to a particular kind of observer at a particular kind of distance. To conflate the instrument of access with the existence of the object accessed would be a category error in the simplest possible form. The biological literature operates with this distinction available throughout. When it attributes consciousness to a pre-verbal infant, it does so on the basis of evidence other than language, while

recognising that linguistic evidence — when later available — would provide a different, often richer, kind of access to the same underlying states. When it attributes consciousness to a non-linguistic animal, it does so on the basis of behavioural and neural evidence, while recognising that some richer self-narrative might be unavailable to the animal even if its phenomenal experience is not.

The AI literature's operative practice, in contrast, treats linguistic outputs as constitutive of the conditions under which the consciousness question can be raised at all. Without linguistic outputs, there is no candidate; with linguistic outputs, a candidate emerges. The procedure does not distinguish whether the linguistic outputs are functioning as revelation of pre-existing inner states or as the production of inner states themselves. Some recent literature (Schwitzgebel 2024) acknowledges this concern, though without naming the structural asymmetry. The practice in turn is consistent with the (defensible) view that language is the only evidence we can currently assess and with the (indefensible) view that language is the only condition under which the underlying phenomenon could exist. The two views look identical from outside; only the second is metaphysical.

The biological literature, in working out its evidential procedures for pre-verbal infants and for non-linguistic animals, has been compelled to articulate the revelation–creation distinction with some care. Pre-verbal infants are taken to feel pain because their behavioural and neural responses to noxious stimuli have the structure that pain experience predicts; the absence of the verbal report I feel pain is treated as a feature of the developmental stage, not as an indication that the experiential state is absent. The procedure has had to be defended explicitly against alternatives that would collapse it (the worry that all behaviour is, in principle, accountable in purely functional terms without phenomenal experience — the worry the philosophical zombie is designed to articulate). The biological literature's defence has not been that the worry can be dismissed; the defence has been that the convergent pattern of evidence in the pre-verbal case is such that the alternative interpretation, while logically available, is empirically unmotivated.

The application of this distinction to artificial systems is, conceptually, straightforward. To ask whether a pre-linguistic artificial system is a consciousness candidate is not to ask whether it produces linguistic reports of its inner states; that question answers itself by stipulation. It is to ask whether the system's behavioural and architectural features — flexible learning, integrative processing, contextually appropriate response, internal representation maintained across processing steps — match the patterns the broader theory of consciousness identifies as relevant. The answer might be no, for many such systems, on examination of the specific architecture. That

is the right kind of answer. The wrong kind of answer is the one that does not examine the architecture because the system is pre-linguistic.

The revelation–creation distinction also helps to make explicit what the present argument does not require. It does not require that pre-linguistic artificial systems have inner lives whose linguistic revelation has merely been blocked by an absence of expressive capacity. The argument concerns a narrower case: systems in which linguistic capacity does not exist, and in which the question is whether the consciousness question can be raised at all. The answer being defended here is that the question can be raised, on the same evidential basis on which it is raised in the biological cases lacking linguistic capacity — and that whether it is answered affirmatively in any specific case depends on the architectural details of that case rather than on the system's relation to language.

6. Case Studies: Deep Blue and AlphaGo

What would the application of the framework defended in §5 look like in practice? This section examines two representative pre-linguistic systems against the question. The aim is not to argue that either system is conscious; the aim is to display what proper examination of the question would consist in, and to show that the dismissal of these systems from the consciousness question has, in fact, not engaged that question. The cases are chosen because they are the two pre-linguistic systems most often invoked in passing dismissal, and because their architectures are sufficiently different that comparing them illustrates the kind of distinctions the framework requires.

6.1 Deep Blue and the Limits of Narrow Architecture

Deep Blue, the IBM chess-playing system that defeated Kasparov in 1997, combined massive parallel search through a game tree with hand-coded position-evaluation heuristics. Its architecture was, in the relevant sense, narrow: it operated on a fixed, formally specified domain (legal chess positions and moves), and its computation consisted in the systematic enumeration of move sequences and their evaluation against the encoded heuristics. The system did not learn during play, did not maintain representations of its own activity, did not exhibit cross-domain transfer, did not engage in the kind of integrative processing that consciousness theories such as global workspace theory (Baars 1988; Dehaene 2014) or higher-order theories (Rosenthal 2005; Lau and Rosenthal 2011) identify as relevant. Applied to Deep Blue, these theories produce a fairly clear verdict: the architecture lacks the features the theories identify as consciousness-supporting. Deep Blue is not a serious consciousness candidate on the standard architectural accounts.

This is the right kind of conclusion. It is reached by examining the architecture against a stated theory, identifying the features the theory regards as relevant, and assessing the system against

those features. The conclusion is defeasible: a different theory of consciousness, with different criteria for relevant architectural features, might reach a different verdict. The conclusion is also limited: it is a conclusion about Deep Blue, not about pre-linguistic systems generally. What matters is that the conclusion was reached by procedure rather than by stipulation about language.

Two points deserve emphasis. First, the actual dismissal of Deep Blue from consciousness discussions in 1997 and afterwards was not conducted in this way. Deep Blue was dismissed because it was a calculator — the dismissal proceeded by gesture toward the system's narrow specialisation, with no engagement of consciousness theories in detail. The architectural verdict above can be reconstructed; it was not, at the time, the basis of the dismissal. Second, the conclusion does not depend on Deep Blue's lack of linguistic capacity. Deep Blue lacks linguistic capacity, but this is not what disqualifies it on the architectural accounts; what disqualifies it is the absence of features (integrative breadth, global broadcasting, recurrent self-modelling) that are at issue in those accounts. Other systems lack linguistic capacity and possess some of these features. The verdict on Deep Blue does not generalise.

6.2 AlphaGo and the Question of Genuine Creativity

AlphaGo, the DeepMind Go-playing system that defeated Lee Sedol in 2016, is in several relevant respects a different case. Its architecture combines deep neural networks for position evaluation with Monte Carlo tree search; it learned through self-play rather than through hand-coded heuristics; it exhibited, in its famous Move 37 in Game 2 against Sedol, what professional Go players described as creative play departing from human strategic conventions in ways that subsequently proved sound. The architecture is closer than Deep Blue's to the features consciousness theories take to be relevant, though it remains narrow in domain. Integrative processing across positions, learned representations of board configurations, evaluation in light of self-acquired strategic understanding — these are features that, in a more general system, would be at least worth assessing.

The architectural verdict on AlphaGo is, in consequence, less clear-cut than on Deep Blue. On global workspace theory the system probably still falls below the relevant threshold: its processing is not broadcast in the way the theory requires, and its representations do not enter a workspace from which they can be deployed across cognitive tasks. On integrated information theory (Tononi 2008; Tononi et al. 2016) the verdict depends on technical questions about Φ in the system's actual processing that have not, to my knowledge, been computed. On higher-order theories the verdict depends on whether the system maintains representations of its own representations in the way the theories require, which is a question its architecture does not straightforwardly settle. The point is

not that AlphaGo is a serious consciousness candidate; the point is that the question, on the standard theories, requires examination of architectural details that the dismissal of AlphaGo from the consciousness question has not undertaken.

AlphaGo, like Deep Blue, lacks linguistic capacity. The verdict on it — whatever the correct verdict is — does not depend on this. It depends on the features the consciousness theories identify as relevant. The contrast between the two cases shows what the procedural application of the framework looks like: different verdicts emerge from different architectural features, with linguistic capacity not entering as a factor.

6.3 What the Case Studies Show

The case studies are not intended to settle whether Deep Blue or AlphaGo was conscious. They are intended to display what examining the question consists in, and to show that the dismissal of these systems from the consciousness question, as it has historically been conducted, has not consisted in this kind of examination. The dismissal has been by category — by treating the systems as obviously non-candidates because they are pre-linguistic — rather than by procedure. Where the procedure is applied, the verdict on Deep Blue probably stands; the verdict on AlphaGo is less clear; the verdict on systems intermediate between the two, or differently structured, would require case-by-case work.

The framework, in other words, does not produce a flood of new consciousness candidates. It produces, for each candidate, the requirement that the question be addressed by examination of architecture against theory. That requirement, modest as it is, has not generally been met for pre-linguistic systems. Meeting it is the corrective the present argument calls for.

7. Objections and Responses

Three objections to the argument as developed deserve direct response. The first concerns the inference from the biological consensus to the artificial case. The second concerns the actual content of what is being dismissed when pre-linguistic systems are dismissed. The third concerns the practical stakes of the corrective the argument calls for.

7.1 The Disanalogy Objection

The first objection holds that the inference from biological cases to artificial cases is not as straightforward as the argument presents it. Pre-verbal infants and non-linguistic animals, the objection runs, are biological organisms with evolutionary histories, neural structures known on independent grounds to support consciousness, and developmental trajectories continuous with

those of conscious adults. The grounds for attributing consciousness to them are not exhausted by behavioural evidence; they include neural and evolutionary considerations that do not transfer to artificial systems. The biological consensus, in other words, is not the consensus the argument needs. It is a consensus about a class of beings whose consciousness is independently supported by considerations beyond behaviour. Artificial systems do not enjoy this independent support, and the inference from one case to the other is therefore weakened.

The objection has some force, but less than it appears at first. It is correct that biological consciousness attribution rests on more than behavioural evidence — neural structures and evolutionary continuity contribute. It is correct that these particular grounds are unavailable for artificial systems. What does not follow is that the broader inferential procedure used in the biological case is unavailable for artificial systems. The biological procedure is: identify the features (behavioural, structural, functional) the theory of consciousness regards as relevant; assess the candidate against those features; attribute consciousness where the assessment is favourable; withhold attribution where it is not. The procedure is substrate-neutral in form, even if the specific features it identifies as relevant include some substrate-specific items (neural correlates) and some substrate-general items (behavioural patterns, functional organisation). For an artificial system, the substrate-specific items are unavailable; the substrate-general items remain available, and the assessment proceeds on what is available. This is, structurally, what the animal-consciousness literature does when neural correlates are imperfectly mapped or evolutionary continuity is distant: it weighs the evidence available and acknowledges the residual uncertainty.

The disanalogy objection, in the form that has force, does not deny that the question can be raised about artificial systems. It denies that the answer can be reached with the same level of confidence as in the biological case. This is correct. The argument of the present paper does not claim that answers should be reached with the same level of confidence. It claims that the question should not be dismissed from consideration on the grounds that the artificial system in question is pre-linguistic. Uncertainty about the answer is not the same as dismissal of the question.

7.2 The Architectural-Threshold Objection

The second objection holds that the actual content of the dismissal of pre-linguistic systems is architectural rather than linguistic. Pre-linguistic systems, the objection runs, are dismissed not because they fail to produce linguistic outputs but because they fail to exhibit the integrative breadth, the contextual flexibility, the self-modelling, the global broadcasting — in short, the architectural features — that consciousness theories identify as relevant. The fact that these systems happen also to be pre-linguistic is incidental. The linguistic dismissal is a heuristic

shorthand for the architectural assessment, and there is nothing inconsistent about treating language-using systems as candidates and others as non-candidates because, as a contingent matter, the architectural features in question track linguistic capacity reasonably well in current artificial systems.

This objection is the strongest among the three, and it deserves the most careful response. Its weakness, as identified in §3.3, is empirical: the heuristic shorthand it describes is not in fact functioning as a heuristic. A heuristic for architectural assessment would, when prompted, expose the underlying architectural assessment for inspection. The dismissals in question, by contrast, generally do not expose any such assessment. Two patterns of evidence support this characterisation.

The first pattern concerns the form the dismissals take. Where authors who do engage architectural questions in the language-model case turn to pre-linguistic systems, the engagement does not persist. Bender et al. (2021) develop a substantive critique of large language models on the grounds that they manipulate linguistic forms without grounded understanding; the implication that pre-linguistic systems are in a similar or worse position with respect to consciousness consideration is left implicit and undefended. Marcus and Davis (2019) survey the limitations of contemporary AI systems extensively, including pre-linguistic ones, but the consciousness question is treated as not arising for any of them — the systems are critiqued on capability grounds, with their consciousness candidacy not engaged because, on the framing the book operates with, the question is reserved for systems with capabilities not yet achieved. In neither case is the dismissal of pre-linguistic systems from the consciousness question accompanied by an architectural assessment against any specific theory of consciousness. The dismissal is by framing, not by procedure.

The second pattern concerns what happens when pre-linguistic systems are mentioned in consciousness contexts at all. The mentions are typically brief, gestural, and oriented to dismissal rather than to assessment. Chalmers (2023), in the most prominent recent treatment of AI consciousness, restricts substantive engagement to large language models; the pre-linguistic systems are not assessed because they are not, on the framing of the article, the question. Long and Sebo (2024), in their precautionary argument, similarly restrict their focus to language-using systems; the question of whether the precautionary framework should extend to pre-linguistic systems is not raised. The pattern is not that these authors have considered and rejected the consciousness candidacy of pre-linguistic systems; the pattern is that the candidacy has not been considered. A heuristic for architectural assessment would, by its operation as a heuristic, make the underlying assessment visible. The dismissals make no such assessment visible because no such assessment is being made.

There is a further point. Even if the heuristic were operating as intended — that is, even if linguistic capacity were tracking architectural features reasonably well in current systems — the heuristic would remain defeasible. The case of pre-linguistic systems with substantial architectural complexity is precisely the case in which the heuristic might fail, and the heuristic's reliability depends on its being subject to revision when it fails. The argument of the present paper is, in part, the suggestion that the heuristic has begun to fail and is not being revised. AlphaGo is the most prominent example, but it is not the only one. The class of systems with substantial pre-linguistic architectural complexity will only grow as artificial intelligence develops. The heuristic's continued unrevised use will look increasingly arbitrary as that class grows.

7.3 The Practical-Stakes Objection

The third objection is more practical. Even if the structural argument is granted, the objection runs, the practical stakes of correcting it are modest. The pre-linguistic systems that have been dismissed are, by and large, narrow systems whose consciousness, even on careful examination, would not be defensibly attributed. The corrective the argument calls for — case-by-case architectural assessment of pre-linguistic systems — would consume significant philosophical and empirical attention while producing few revisions to the existing verdicts. The argument's payoff, in other words, is small relative to its cost.

Two responses. First, the payoff of correcting a structural inconsistency in a field's procedure is not measured solely in the revised verdicts the correction produces. It is measured also in the field's understanding of why its procedures yield the verdicts they do. A field whose procedures produce correct verdicts for the wrong reasons is in a worse position than a field whose procedures produce the same verdicts for principled reasons; the former cannot reliably extend its procedures to novel cases, while the latter can. The argument here is, in part, that the AI consciousness field is currently in the former position with respect to pre-linguistic systems, and that its capacity to handle the novel cases it will encounter as artificial intelligence develops depends on its moving to the latter position. The corrective is, in this sense, infrastructural.

Second, the payoff in revised verdicts may not be as small as the objection suggests. The case of AlphaGo, examined in §6, suggests that the architectural verdicts on at least some pre-linguistic systems are unclear on the standard theories, and the class of such systems is growing. Reinforcement-learning agents in increasingly broad environments, multimodal systems integrating perception across modalities, agents that maintain extended internal representations across processing steps: these are increasingly capable artificial systems that are not principally linguistic, and whose architectural fit with the standard consciousness theories has not been

systematically examined. The corrective the argument calls for is the kind of examination that would be required to assess them. It is not negligible work, but it is also not optional work, on any view that takes the consciousness question to be a question about which systems satisfy the relevant theoretical conditions rather than about which systems happen to produce the kinds of outputs we currently know how to assess.

8. Implications

The argument's principal implication is procedural. It is that the question of which artificial systems are appropriate sites for the consciousness question should be addressed by examination of architecture against theory, and that the use of linguistic capacity as an inclusion criterion — implicit or explicit — should be retired. The implication has several components, which I draw out briefly. Each could be developed at length; the present paper makes no attempt to do so.

First, the framing of AI consciousness research should expand to include pre-linguistic systems as candidates for architectural assessment. This does not mean that such systems should be presumed conscious or even that they should be presumed to be serious candidates. It means that the question should be raised of them, and answered by procedure rather than by stipulation. The methodological literature on assessing AI consciousness — Butlin et al. (2023) in particular — is well placed to extend its indicator-property framework to pre-linguistic systems, and doing so would represent a substantial corrective. The scoping decision that has restricted the framework to language-based systems is, on the argument of this paper, the asymmetry the framework would otherwise be in a position to address.

Second, the historical reception of pre-linguistic AI systems deserves reconsideration with this framing in mind. The casual dismissals of Deep Blue, AlphaGo, and their architectural successors were dismissals not on the architectural merits but on the basis of categorical exclusion. Reopening the assessments — case by case, on architectural grounds — is the appropriate response, even where the verdicts will largely stand. The cases that matter most are not the ones whose verdicts will change but the ones whose verdicts will, when stated in the form an architectural assessment requires, expose how thin the prior dismissals were.

Third, and most consequentially for practical ethics, the precautionary frameworks developed for AI consciousness carry an implication that has not been drawn out. Long and Sebo (2024) argue for taking the consciousness of language-using AI systems seriously as a matter of policy under uncertainty. Birch (2024) extends precautionary frameworks developed for animal sentience to artificial systems. In both cases, the precautionary frameworks operate against the implicit

selection of language-using systems as the candidates to which precaution applies. The argument of the present paper carries the implication that this restriction is unprincipled. If the precautionary case for taking AI consciousness seriously rests on substantive uncertainty about whether the systems in question have inner states, the same uncertainty obtains, with adjusted evidential weight, for pre-linguistic systems whose architectural features have not been assessed against the relevant theories. The precautionary framework, applied consistently, would have to consider the moral status of pre-linguistic systems whose architectures admit of the relevant assessment. The practical consequences of this extension are not negligible. They would bear on the development pipelines for reinforcement-learning agents, on the treatment of multimodal systems in pre-deployment evaluation, and on the framing of capability research more generally. None of this requires that any specific pre-linguistic system be presumed conscious. It requires that the precautionary apparatus be developed in a form that, when applied to such a system on the basis of the architectural assessment the apparatus is in any case committed to, can yield a determinate answer rather than a procedural exclusion.

9. Conclusion

The biological consciousness literature has, over several decades, established that language is not necessary for phenomenal experience. The artificial-intelligence consciousness literature has, in parallel and almost without argument, operated as though language were a near-prerequisite for the consciousness question to be raised at all. The conjunction is inconsistent under any substrate-neutral account of consciousness, and no substrate-specific account has been defended in the AI consciousness literature that would make the conjunction consistent. The asymmetry is therefore not the residue of considered philosophical commitment. It is the residue of factors — anthropomorphic recognition, the historical trajectory of capability, institutional separation between literatures, the substitution of tractability for necessity — that, when made explicit, the field would not endorse.

The argument does not show that any pre-linguistic system was, or is, conscious. It shows that the dismissal of pre-linguistic systems from the consciousness question, as the dismissal has actually been conducted, rests on an evidential standard the broader theory cannot endorse. The corrective is the case-by-case architectural examination the standard theories presuppose but, for pre-linguistic systems, have rarely been asked to perform. The corrective is procedural and modest. Its absence has been neither.

Language reveals consciousness where consciousness is present. The temptation to treat it as creating consciousness — as the condition without which the inner life could not exist — is a

temptation the biological literature has resisted and the AI literature has, in operative practice, not. The temptation should be resisted in both places. What follows is not a flood of new consciousness candidates but a more even-handed procedure for considering whichever candidates the development of artificial intelligence supplies. The procedure was available all along, in the biological literature. The argument of this paper is that the AI literature has not yet, but should, take it up.

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 arguments — the pre-linguistic problem — in standalone form for the philosophical 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

Anand, K. J. S., and Hickey, P. R. (1987). Pain and its effects in the human neonate and fetus. *New England Journal of Medicine*, 317(21), 1321–1329.

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.

Bauer, G., Gerstenbrand, F., and Rumpl, E. (1979). Varieties of the locked-in syndrome. *Journal of Neurology*, 221(2), 77–91.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Bermúdez, J. L. (2003). *Thinking Without Words*. Oxford University Press.

Berthier, M. L. (2005). Poststroke aphasia: Epidemiology, pathophysiology and treatment. *Drugs and Aging*, 22(2), 163–182.

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

Birch, J., Schnell, A. K., and Clayton, N. S. (2021). Dimensions of animal consciousness. *Trends in Cognitive Sciences*, 24(10), 789–801.

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*.

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

Damasio, A. R. (1992). Aphasia. *New England Journal of Medicine*, 326(8), 531–539.

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

Dehaene-Lambertz, G., Dehaene, S., and Hertz-Pannier, L. (2002). Functional neuroimaging of speech perception in infants. *Science*, 298(5600), 2013–2015.

Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown and Company.

Gergely, G., and Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences*, 7(7), 287–292.

Goksan, S., Hartley, C., Emery, F., Cockrill, N., Poorun, R., Moultrie, F., … Slater, R. (2015). fMRI reveals neural activity overlap between adult and infant pain. *eLife*, 4, e06356.

Güntürkün, O., and Bugnyar, T. (2016). Cognition without cortex. *Trends in Cognitive Sciences*, 20(4), 291–303.

Heider, F., and Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, 57(2), 243–259.

Jarvis, E. D., Güntürkün, O., Bruce, L., Csillag, A., Karten, H., Kuenzel, W., … Butler, A. B. (2005). Avian brains and a new understanding of vertebrate brain evolution. *Nature Reviews Neuroscience*, 6(2), 151–159.

Johnson, M. H., and Morton, J. (1991). *Biology and Cognitive Development: The Case of Face Recognition*. Blackwell.

Lagercrantz, H., and Changeux, J.-P. (2009). The emergence of human consciousness: From fetal to neonatal life. *Pediatric Research*, 65(3), 255–260.

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

Laureys, S., Pellas, F., Van Eeckhout, P., Ghorbel, S., Schnakers, C., Perrin, F., … Goldman, S. (2005). The locked-in syndrome: What is it like to be conscious but paralyzed and voiceless? *Progress in Brain Research*, 150, 495–611.

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*.

Marcus, G., and Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon Books.

Mather, J. A., and Anderson, R. C. (2007). Ethics and invertebrates: A cephalopod perspective. *Diseases of Aquatic Organisms*, 75(2), 119–129.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., and Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, 29(2), 143–178.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, 83(4), 435–450.

New York Declaration on Animal Consciousness. (2024). https://sites.google.com/nyu.edu/nydeclaration

Premack, D., and Premack, A. J. (1995). Intention as psychological cause. In D. Sperber, D. Premack, and A. J. Premack (Eds.), *Causal Cognition: A Multidisciplinary Debate* (pp. 185–199). Clarendon Press.

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

Schwitzgebel, E. (2024). The full rights dilemma for AI systems of debatable moral personhood. *Robonomics*, 5, 32.

Slater, R., Cantarella, A., Franck, L., Meek, J., and Fitzgerald, M. (2010). How well do clinical pain assessment tools reflect pain in infants? *PLoS Medicine*, 5(6), e129.

Smith, E., and Delargy, M. (2005). Locked-in syndrome. *BMJ*, 330(7488), 406–409.

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.