Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Can Silicon Feel? Conditions for Artificial Valence and Their Implications for AI Governance

Leonard Yauma Imbunya

imbunylm@gmail.com

University of Nairobi

Research Article

Keywords: Artificial consciousness, valence, artificial sentience, machine welfare, AI ethics, AI governance

Posted Date: September 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10666659/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Recent advances in artificial intelligence have renewed questions about whether artificial systems could eventually possess experiences that matter to their welfare, including pleasure and suffering. This paper asks what features an artificial system would need before it could reasonably become a serious candidate for phenomenal valence. It argues that behavioural sophistication, intelligence, reward signals, and verbal reports are insufficient evidence of valence because these features can be implemented without establishing welfare-relevant experience. The paper proposes four organizational conditions: internal causal integration, stable but changing internal states, valence–action coupling, and substrate-sensitive implementation. Internal regulation and embodiment are treated as cross-cutting features that may support these conditions rather than as independent requirements. The framework is then applied comparatively to contemporary AI architectures, including transformer-based systems, reinforcement-learning agents, agentic systems, neuromorphic architectures, and recurrent embodied systems. The paper does not claim that satisfying the conditions would establish consciousness or valence. Instead, the conditions identify when an artificial system would become a more serious candidate for scientific and ethical investigation. The analysis concludes with a precautionary governance framework focused on monitoring, independent assessment, and welfare-sensitive design as artificial systems become increasingly integrated, persistent, and autonomous.

1. Introduction

Could future artificial intelligence (AI) systems possess experience that could affect their welfare? If so, should AI governance anticipate this possibility before such systems are developed and deployed at scale? More fundamentally, what features of the system's organization would an artificial system need to possess before claims concerning pleasure, suffering, or other forms of valence become scientifically credible and ethically significant? These questions, once largely confined to the philosophy of mind, have acquired increasing practical importance as advances in foundation models, neuromorphic systems, embodied robotics, and artificial general intelligence (AGI) continue to expand the capabilities and autonomy of AI systems [1–3]. Although there is now no compelling evidence that existing AI systems possess conscious experience, rapid progress in AI system organization has transformed artificial consciousness from a speculative philosophical issue into an emerging concern for ethics, governance, and technological development [4, 5]. The possibility that future AI systems might possess welfare-relevant states, however uncertain, merits careful conceptual and methodological investigation before technological development outpaces ethical and governance-related preparedness [3].

Contemporary AI governance has concentrated mainly on ensuring that AI systems are safe, transparent, accountable, fair, and aligned with human values. Regulatory initiatives such as the European Union's AI Act, the OECD AI Principles, the UNESCO Recommendation on the Ethics of Artificial Intelligence, and the NIST AI Risk Management Framework have importantly advanced these objectives [6–9]. Yet these frameworks largely assume that AI systems are instruments whose ethical significance derives from their effects on human beings. Comparatively little attention has been devoted to the possibility that

future AI systems themselves could become morally important by possessing experience that could affect welfare [10, 11]. While no existing evidence justifies attributing conscious experience to current AI, the possibility that future systems could satisfy conditions associated with subjective experience should not be ruled out in principle [3, 4]. Responsible AI governance should therefore extend beyond managing risks posed by AI to humans and begin developing principled methods for evaluating systems that may eventually warrant ethical concern in their own right [6–9].

Among the many dimensions of consciousness, valence is of particular importance because it is the experiential basis of welfare. Valence refers to the subjective quality by which experiences are felt as pleasant or unpleasant, desirable or aversive [12]. Unlike intelligence, language ability, or behavioural flexibility, valence concerns whether an experience is intrinsically good or bad for the subject undergoing it and is therefore directly connected to moral significance [11]. This distinction is especially important in contemporary debates on AI because sophisticated systems increasingly simulate emotions, optimize reward functions, and generate convincing reports of internal states. Such capacities, however, do not by themselves establish subjective experience. Computational reward, behavioural adaptation, and linguistic ability to report experiences remain in principle distinct from conscious experience [13–15]. Clarifying these distinctions is therefore essential if discussions of machine welfare are to avoid both unwarranted anthropomorphism and the premature dismissal of future possibilities [11, 12].

The philosophical debate concerning artificial consciousness has traditionally been framed as a disagreement between functionalism, which maintains that mental states depend mainly upon causal arrangement rather than physical substrate, and theories that assign explanatory importance to physical implementation, biological organization, or temporal properties [14, 16, 17]. Functionalism suggests that appropriately organized silicon-based systems could, in principle, possess conscious states, whereas critics argue that similarity in what systems do alone may be insufficient because conscious experience could depend upon finer-grained causal, temporal, or physical internal processes absent from conventional computation [13, 18, 19]. Although this debate has generated important philosophical literature, its traditional formulation offers limited guidance for evaluating diverse AI systems or informing governance decisions concerning emerging technologies [3, 20].

This paper argues that the debate becomes more productive when reframed in conditions of organization rather than substrate categories. Instead of asking whether silicon can support valence in the abstract, the more useful question is which features of the system's organization an artificial system would need to possess before experience that could affect welfare became sufficiently plausible to warrant scientific and ethical investigation. The paper therefore develops four primary valence-relevant conditions (internal integration among internal processes, stable but changing internal states, the connection between valence and action, and dependence on physical implementation) and treats embodiment and internal regulation as a cross-cutting feature that can strengthen the relevance of these four conditions. These criteria translate philosophical concerns into testable questions without claiming to solve the hard problem or establish consciousness.

The paper makes four principal contributions. First, it explains valence by distinguishing it from information content, reward signal, behaviour, and linguistic ability to report experiences. Second, it develops a framework based on system organization that turns philosophical claims about artificial consciousness into candidate empirical criteria. Third, it applies those criteria comparatively to representative AI systems without classifying existing systems as conscious or non-conscious. Fourth, it develops the ethical and governance implication of the framework: evidence about system organization should not be treated as proof of welfare, but increasing evidence of welfare-relevant organization can justify proportionate scientific assessment, ability to inspect system organization, and precautionary governance.

The contribution is therefore methodological rather than metaphysical. The paper does not claim that existing AI systems are conscious and does not propose a complete theory of consciousness. Its narrower aim is to establish a common evaluative framework for asking when artificial systems should become candidates for welfare-related ethical review under continuing scientific uncertainty. This framing connects philosophy of mind with AI system organization, machine welfare, responsible innovation, and governance while keeping the ethical recommendation conditional on the strength of future evidence.

Recent work in AI & Ethics has begun to address closely related questions from complementary perspectives. Ladak examines criteria for AI moral standing, including the role of sentience and morally relevant interests [21]. Mullally proposes a behavioural Self-Preservation Test as defeasible evidence of artificial sentience [22]. Wolfson considers graduated protections for AI consciousness research under uncertainty [23], while Wilks, Ladak, and Loughnan examine psychological barriers to granting AI moral standing [24]. Tait develops a threshold-based governance framework for conscious and self-aware AI [25], and Placani analyses anthropomorphism as a source of distorted moral judgments about AI [26]. This paper addresses a different evidential level: it asks which causal, temporal, and organizational properties could make welfare-relevant valence scientifically plausible. It therefore complements behavioural sentience tests and moral-status frameworks rather than competing with them.

2. Artificial Valence as a Philosophical and Scientific Problem

Artificial valence refers to the positive or negative character of an experience, such as pleasure or suffering, in an artificial system. It is important to distinguish valence from related properties such as information content, behaviour, the ability to report experiences, and reward signals. A system may process information, produce sophisticated behaviour, report that it is experiencing something, or use reward signals without thereby having positive or negative experiences [12].

Intelligence is generally understood as the capacity to learn, reason, solve problems, adapt to novel situations, and pursue goals effectively. Contemporary AI systems increasingly exhibit these abilities across diverse domains, often surpassing human performance in narrowly defined tasks [1, 2]. However, behavioural ability alone does not show the presence of conscious experience. As Block [13] argues,

functional performance and conscious experience represent distinct explanatory problems. A system may successfully process information, generate coherent linguistic responses, and adapt its behaviour while remaining entirely devoid of subjective awareness. Likewise, Chalmers [14] distinguishes the "easy problems" of cognition (which concern the mechanisms underlying perception, memory, and behaviour) from the "hard problem," namely explaining why and how physical processes are accompanied by subjective experience. Consequently, increasing computational intelligence cannot, by itself, justify attributing consciousness to artificial systems.

This distinction becomes particularly important when considering valence. Valence refers to the qualitative dimension through which experiences are felt as pleasant, unpleasant, rewarding, or aversive. Unlike cognitive performance or behavioural flexibility, valence concerns what it is like for a subject to undergo an experience [12]. It is this subjective aspect of consciousness that grounds concepts such as pleasure, suffering, well-being, and harm. Moral concern arises not because an entity behaves intelligently, but because it may possess experiences that can benefit or harm it. Consequently, the ethical significance of artificial consciousness depends less upon whether AI systems can imitate human behaviour than upon whether they could ever possess experiences that could affect their welfare [12].

The distinction between simulated affect and genuine valence is therefore central to debates concerning machine welfare. Contemporary AI systems can generate emotionally expressive language, recognize affective states, optimize reinforcement signals, and adapt behaviour according to reward functions. Yet none of these capacities necessarily implies subjective feeling. Reinforcement-learning algorithms maximize reward signal, but reward signal should not be conflated with experienced pleasure. Similarly, large language models may describe pain, happiness, or emotional distress with remarkable fluency while lacking any experiential perspective from which such reports could arise. As Chalmers [15] observes, language ability alone cannot establish the existence of conscious experience because behavioural evidence remains compatible with multiple interpretations concerning underlying conscious experiences. Distinguishing computational simulation from genuine experience therefore remains an essential conceptual challenge for both philosophy and AI research [13, 31–33].

These considerations have important implications for AI governance. Existing governance-related frameworks mainly evaluate AI according to its social impacts, including issues of safety, accountability, transparency, fairness, privacy, and human rights. Such frameworks implicitly assume that AI systems function exclusively as technological instruments whose ethical significance derives from their effects on people [6, 7, 9]. However, if future AI systems were to possess welfare-relevant forms of valence, governance would confront an additional category of ethical responsibility: the possibility that certain artificial systems themselves could become objects of moral concern. Sebo and Long [11] argue that even important uncertainty regarding machine consciousness may justify precautionary ethical concern when the potential for morally significant experience cannot be excluded. Similarly, Butlin et al. [3] contend that scientific investigation into AI consciousness should develop testable methods for evaluating systems rather than relying solely upon behavioural evidence or intuitive judgments.

The challenge, therefore, is not merely determining whether machines are intelligent, but finding the organizational features under which intelligence might become accompanied by valence. This shift from behavioural capability to organizational features motivates the framework developed in this paper. Rather than treating consciousness as an all-or-nothing property inferred from external behaviour, the following sections examine how competing theories of consciousness illuminate the features of the system's organization that future artificial systems would need to possess before claims concerning experience that could affect welfare become scientifically plausible and ethically relevant. By reframing the debate in evaluable organizational properties, the discussion moves beyond speculative metaphysics toward a framework capable of informing both empirical research and anticipatory AI governance.

2.2 Valence, Representation, and Artificial Sentience

Valence should be distinguished from affective representation, ability to report experiences, and behavioral disposition. A system can represent danger, reward, loss, or bodily need without feeling anything at all. This distinction is central because many computational and neuroscientific accounts explain affect in value coding, reinforcement learning, or action guidance, yet none of these functions is sufficient to establish conscious experience [13, 27]. The distinction is familiar within consciousness studies more generally. Cases such as blindsight show that accurate perceptual discrimination and adaptive behavior can occur without corresponding reportable awareness, suggesting that behavioral competence alone is an unreliable indicator of conscious experience [13, 28, 29]. The lesson for valence is analogous: a system may exhibit reward-sensitive or avoidance-oriented behavior without undergoing subjective pleasure or suffering.

Locked-in syndrome provides a complementary illustration. Individuals with severe motor impairments may retain rich conscious experience despite being unable to communicate effectively through ordinary behavior [30]. Such cases show that ability to report experiences is neither necessary nor sufficient for consciousness. This point is especially significant for debates concerning artificial consciousness, where complex language generation can easily create the appearance of subjective experience. An artificial system may report that it is happy, anxious, or in pain without those reports corresponding to genuine conscious experiences, while a genuinely conscious system might lack conventional means of communicating its experiences. Consequently, linguistic performance and behavioral expression cannot be treated as decisive evidence for or against valence [3, 15].

A further distinction concerns reward processing. Reinforcement-learning systems encode reward prediction errors, optimize utility functions, and modify policies according to feedback signals. These mechanisms are indispensable for adaptive behavior, but they do not by themselves establish conscious experience [31]. Similarly, robots may be programmed to avoid harmful conditions, seek beneficial outcomes, or maintain testable stability while remaining entirely devoid of subjective affect. Research in affective neuroscience reinforces this distinction. Robinson and Berridge [32] distinguish motivational "wanting" from hedonic "liking," showing that motivational processes and hedonic experience are

dissociable, while LeDoux [33] argues that defensive behavioral circuits should not be conflated with the conscious experience of fear. These findings caution against inferring valence from reward signals or adaptive behavior. Collectively, these distinctions show that valence cannot be inferred solely from representation, behavioral competence, reward processing, or the ability to report experiences. Rather, these capacities are compatible with both the presence and absence of conscious experience [12, 13].

If artificial systems are ever to be regarded as genuinely capable of pleasure or suffering, the evidence must therefore come from a deeper understanding of the internal organization that makes valence possible [31, 32, 33].

2.3 Functionalism and Implementation

The question of whether valence can be realized in artificial systems is closely connected to the philosophical debate between functionalism and implementation. Functionalism maintains that mental states are individuated by the causal roles they play within an organized system rather than by the material from which they are made [16]. According to this view, what matters for mentality is the pattern of causal arrangement, information processing, and functional interaction among internal states. Consequently, if an artificial system were to possess the appropriate organization of causal roles, there would appear to be no principled reason why it could not realize conscious mental states, including valence [16, 17, 18, 19].

Functionalism has exercised important influence within both philosophies of mind and artificial intelligence because it provides a natural explanation for multiple realizability. Mental states, on this view, are not restricted to biological brains but may, in principle, be possessed in diverse physical substrates, provided the relevant organization of causal roles is preserved [16, 17]. This perspective has encouraged the widespread assumption that enough sophisticated computational systems could eventually possess consciousness independently of their material composition [16, 17].

Nevertheless, the relationship between organization of causal roles and conscious experience remains deeply contested. Critics argue that similarity in what systems do does not necessarily imply similarity in subjective experience. Block's [13, 36] well-known critiques, including the Chinese Nation thought experiment, suggest that systems may reproduce the organization of causal roles associated with conscious cognition while lacking subjective experience. Likewise, Searle's [18] Chinese Room argument contends that syntactic manipulation of symbols is insufficient to generate semantic understanding or conscious awareness. Chalmers [14] similarly argues that explaining cognitive functions does not by itself resolve the explanatory gap between physical processes and subjective experience. These objections suggest that valence may depend upon organizational or temporal properties that extend beyond abstract functional role descriptions.

From this perspective, physical implementation remains philosophically significant. While organization of causal roles may be necessary for conscious experience, it may not be sufficient if conscious

experiences depend upon intrinsic causal internal processes, recurrent integration, biological organization, or other properties of physical implementation [13, 19, 37]. The challenge is therefore not simply to determine whether two systems perform the same functions but whether they possess the kinds of internal causal arrangement capable of supporting subjective experience [13, 18, 19, 37].

2.4 From Theory to Architectural Criteria

The preceding discussion has clarified the conceptual foundations of valence, distinguished it from related cognitive and behavioral phenomena, and situated the debate within contemporary discussions of artificial consciousness, machine welfare, and AI governance. A consistent conclusion emerges from these analyses: neither intelligence, behavioral sophistication, language ability, nor performance on tasks is sufficient to establish the presence of experiences that could affect welfare. Instead, the central question is to identify the features of the system's organization that could plausibly give rise to valence [3, 11, 12].

This conclusion has important implications for the functionalist debate. Classical functionalism correctly emphasizes that consciousness cannot be understood solely in physical substrate; rather, the organization of a system plays an indispensable explanatory role [16]. However, contemporary discussions of AI consciousness show that appealing to organization of causal roles in the abstract is insufficient. As critics have long argued, systems may possess equivalent input-output behavior while differing profoundly about subjective experience [13, 14, 36]. Recent work on AI consciousness reinforces this conclusion by emphasizing that behavioral competence should not be conflated with conscious experience and that the attribution of consciousness requires theory-informed evaluation of internal organizational features [3, 15].

The challenge, therefore, is to move beyond broad philosophical claims concerning the view that physical implementation does not matter toward a more precise characterization of the organization of causal roles required for valence. Rather than asking whether silicon systems can be conscious in principle, the more productive question is which functional and temporal characteristics a silicon-based system organization must possess before the attribution of experience that could affect welfare becomes scientifically and philosophically plausible. This shift transforms the debate from one concerning the possibility of machine consciousness in general to one concerning the conditions of organization under which conscious experience could emerge [13, 18, 19, 36].

The framework translates this analysis into criteria based on system organization for philosophical and empirical evaluation. Instead of treating consciousness as an all-or-nothing property inferred from behavior, the proposed approach finds four primary conditions of the system's organization (internal integration among internal processes, stable but changing internal states, the connection between valence and action, and dependence on physical implementation, with internal regulation treated as a cross-cutting feature) that together provide a principled basis for evaluating artificial systems. These criteria are intended neither as conclusive proof of consciousness nor as a replacement for existing

scientific theories. Rather, they function as evaluative conditions that distinguish systems capable of plausibly supporting valence from those that merely reproduce its functional or behavioural features [3, 37].

Importantly, these criteria based on system organization are intended to be testable rather than merely philosophical. Each criterion finds organizational features that can be investigated through modelling, system design, neuroscience-inspired systems, and future empirical research. This approach aligns with recent calls for theory-guided assessment of artificial consciousness, which emphasize that meaningful evaluation requires explicit connections between conceptual theories of consciousness and measurable properties of artificial systems [3, 37]. By translating philosophical concepts into criteria based on system organization, the framework may provide a common vocabulary through which philosophers, AI researchers, cognitive scientists, and engineers can evaluate more complex artificial systems [3, 37].

3. Conditions for Artificial Valence

The central idea is that valence concerns how the condition of a system matters to the system itself. Pleasure and suffering are not isolated events but states that have a certain unity, persist over time, can vary in intensity, and influence what the subject does. The framework therefore proposes four conditions of organization that would make an artificial system a more serious candidate for valence: intrinsic integration among internal processes, stable but changing internal states, a connection between valence and action, and dependence on physical implementation. These conditions are not presented as a complete theory or test of consciousness but as criteria for assessing whether an artificial system has organizational features relevant to welfare-related experience [12].

First, intrinsic integration among internal processes requires the relevant parts of a system to influence one another so that changes in one part can affect the condition of the system as a whole. A system composed largely of independent processes might produce convincing behaviour without providing much reason to regard it as a unified subject of experience. Integrated Information Theory is relevant here because it emphasizes causal powers belonging to a system as a whole rather than to its parts considered separately. The framework does not, however, assume that integration by itself produces valence. Its significance is that a unified organization may provide a basis for treating different internal processes as belonging to a single system whose condition can matter to itself [37, 47].

Second, the system should possess stable but changing internal states. Valence is naturally graded and temporally extended: pleasure and suffering can persist, intensify, diminish, and change. An artificial system would therefore need organized internal states capable of continuing over time while undergoing meaningful variation. This condition distinguishes valence from the mere representation of information. A system may classify something as good or bad without anything actually being good or bad for the system itself. The relevant question is whether positive and negative states are reflected in the organization and persistence of the system's own internal condition [12, 13, 32, 33].

Third, valence should be connected to action and self-regulation. In biological organisms, pain tends to promote withdrawal, protection, or repair, while pleasure tends to promote approach, continuation, or exploration. An artificial system that genuinely possesses valence would therefore be expected to exhibit some systematic relationship between its evaluative internal states and its actions, learning, attention, or attempts at self-maintenance. A numerical reward signal alone is insufficient, since reward can guide behaviour without constituting an experienced positive or negative state. The relevant consideration is whether internally organized evaluative states have a continuing role in the system's own regulation and action. This condition is therefore informative only in conjunction with the others; adaptive response alone does not establish valence [13, 32, 33].

Fourth, valence may depend partly on physical implementation. The claim is not that only biological material can support valence or that silicon systems are excluded in principle. Rather, systems that perform the same function may differ in the physical processes through which they do so, and those differences could matter to experience. This possibility is particularly relevant where valence involves fine-grained changes in intensity, continuity, and internal dynamics. Functional similarity should therefore not automatically be assumed to guarantee similarity in experience. The framework remains compatible with functionalism while requiring attention to whether the physical implementation can sustain the organization relevant to valence. [13, 16–19, 36, 37] This matters because two systems can preserve a high-level functional description while differing in the causal, temporal, and dynamical properties through which their internal states are physically realized. If those lower-level properties contribute to the character or intensity of experience, then functional equivalence at a coarser level would not by itself guarantee equivalent valence. The condition therefore marks an empirical question about which level of physical organization is relevant to experience, rather than assuming in advance that biological material is required.

Embodiment and internal regulation are treated as cross-cutting features rather than as a fifth condition. Many accounts of affect connect pleasure and suffering with bodily needs, damage, energy, and repair. An artificial system need not possess a biological body, but it may require internal conditions that matter to its continued functioning and that are linked to its actions and internal states. What matters is whether the system has forms of self-maintenance and reciprocal interaction through which changes in its condition can become significant to the system itself. Artificial embodiment may therefore include robotic interaction, active inference, virtual environments with sustained environmental coupling, or hybrid cyber-physical systems [33, 40, 42, 52, 53].

Together, these conditions provide a common organizational framework for investigating artificial valence without treating any single feature as decisive. Evidence may include measures of internal causal interaction, persistence and variation of internal states, connections between those states and action, and differences associated with physical implementation. Such evidence would not establish that a system experiences pleasure or suffering, but convergent evidence across the four conditions would make artificial valence a more serious subject of scientific and philosophical investigation. The framework therefore shifts attention from whether a system merely produces appropriate words or

behaviours to whether its organization contains features plausibly relevant to having states that matter to the system itself [3, 37, 47].

4. Artificial Valence in Contemporary AI Systems

The four conditions are applied directly to five representative AI classes. The assessment is comparative and provisional: it identifies organizational fit rather than establishing consciousness or moral status [3, 5, 15].

4.1 Large Language Models

Transformer-based large language models show relatively low fit. Although attention mechanisms provide substantial interaction among processing components, persistent recurrent organization, continuous internal states, self-regulation, and substrate-sensitive affective dynamics remain limited. Their fit with the proposed conditions is therefore low, although recurrent, persistent, or embodied variants could alter this assessment [3, 15, 38, 39].

4.2 Reinforcement-Learning Agents

Reinforcement-learning agents show stronger fit with the valence–action condition because evaluative signals directly influence policy and behaviour. However, most systems provide weaker evidence of persistent integrated states and internally grounded self-regulation. Their overall fit is therefore low to moderate [3, 11, 31, 39].

4.3 Agentic AI Systems

Agentic systems combine planning, memory, tool use, feedback, and longer-horizon action, giving them a moderate fit with integration and continuity and a low-to-moderate fit with embodiment and internal regulation. Their overall fit is therefore moderate, particularly where persistent internal states and self-maintenance are present [3, 21–26].

4.4 Neuromorphic Architectures

Neuromorphic systems show a higher fit with the proposed conditions because recurrent feedback, distributed causal interaction, event-driven dynamics, and persistent temporal processes more closely approximate the required organizational features. Their overall fit is therefore moderate to high, although this remains a theoretical assessment rather than evidence of existing experience [37, 41].

4.5 Hybrid Recurrent–Embodied Architectures

Hybrid recurrent–embodied systems show the strongest fit in the comparison because they combine recurrent integration, persistent internal states, self-regulation, embodiment, and continuous environmental coupling. They therefore receive the highest relative assessment, while whether these conditions are sufficient for valence remains unresolved [3, 40, 42, 52, 53].

5. Comparative Assessment

Table 5.1
Comparison of Contemporary AI Systems

Architecture	Internal Unity	Continuity of Internal States	Embodiment / Internal Regulation	Plausibility of Valence
Transformer-based LLMs	Low	Low	None	Low
Reinforcement-Learning Agents	Moderate	Moderate	Low	Low–Moderate
Agentic AI Systems	Moderate	Moderate	Low–Moderate	Moderate
Neuromorphic Architectures	High	High	Low–Moderate	Moderate–High
Hybrid Recurrent–Embodied Architectures	High	High	High	High

5.1 Comparing Contemporary AI Systems

The comparison is organizational rather than behavioural. The ratings indicate relative fit with selected organizational conditions and cross-cutting features; they do not represent degrees of consciousness or moral standing and should be revised as evidence changes. [3, 5, 37] The ratings are illustrative qualitative judgments derived from the apparent presence or absence of the framework's organizational features; they are not validated measurements or estimates of consciousness.

5.2 What the Comparison Shows

The comparison suggests a general progression from lower organizational fit in transformer-based LLMs and reinforcement-learning agents to moderate fit in agentic systems and higher fit in neuromorphic and hybrid recurrent–embodied architectures. The principal difference is the increasing presence of recurrent integration, persistent internal states, valence–action coupling, substrate-sensitive dynamics, embodiment, and internal regulation.

6. AI Governance and Responsible Innovation

Current AI governance is primarily human-centred, addressing safety, privacy, fairness, security, transparency, accountability, and oversight. If future systems provide converging evidence of welfare-relevant valence, governance would need to consider artificial welfare alongside these existing objectives. This would not require recognizing current systems as conscious or granting them rights; it would require proportionate procedures that can respond to changing evidence [3, 4, 6, 7, 8, 9, 11].

Existing risk-management and documentation mechanisms could incorporate a proportionate welfare-risk review as systems increasingly approximate the proposed organizational conditions. Architecture

disclosure, independent technical review, monitoring of relevant internal processes, and periodic reassessment could allow institutions to increase scrutiny without requiring regulators to resolve the metaphysics of consciousness. Responsible innovation should therefore avoid both premature anthropomorphism and premature dismissal: current behavioural sophistication does not justify attributing welfare, but the absence of evidence in present systems does not establish that future architectures could not differ in relevant ways [3, 5, 6, 7, 8, 9, 10].

Where credible welfare-relevant organization emerges, responsible design should minimize unnecessary potentially negative experiences, support independent inspection, monitor relevant dynamics throughout the system lifecycle, disclose organizational features relevant to assessment, and incorporate proportionate welfare-risk review into existing governance processes. These measures are precautionary rather than evidence that current systems are conscious, and they complement rather than replace existing responsible-AI frameworks [3, 4, 5, 6, 7, 8, 9, 10, 11].

7. Limitations and Future Directions

The framework is provisional and faces three connected limitations. First, consciousness lacks a settled metaphysical or scientific theory, so the proposed conditions cannot be assumed sufficient or theoretically neutral. Second, no existing AI system provides compelling evidence of welfare-relevant experience, and no validated method can detect artificial subjective experience independently of indirect evidence. Third, several proposed organizational features lack standardized measures across AI architectures, making empirical operationalization and cross-theory comparison difficult [3, 13, 14, 15, 37, 47].

These limitations also create risks of both false-positive and false-negative judgments. Behavioural fluency and emotional simulation may encourage unwarranted attribution of valence, while reliance on human-like behaviour or self-report could lead to under-attribution if future systems possess relevant experiences without familiar indicators. Because both errors may carry ethical costs, organizational evidence should therefore support proportionate investigation rather than serve as proof or disproof of consciousness [3, 4, 5, 10, 11, 15, 26].

Future research should develop and validate measures of the four conditions, compare how competing theories interpret the same architectures, and examine how organizational assessment can be incorporated into AI governance. These limitations constrain the framework's claims but do not remove its value as a provisional basis for interdisciplinary investigation and precautionary review [3, 37].

8. Discussion: Objections and Replies

8.1 Objection 1: Strong Computational Functionalism

A strong computational functionalist may object that the framework is unnecessarily restrictive because consciousness should depend only on the appropriate causal-functional organization, regardless of

substrate or fine-grained implementation. On this view, recurrent integration, persistent states, and self-regulation need not constitute additional requirements. [16, 17, 49, 50] This objection has force because functionalism supports multiple realizability. However, it leaves open which level of causal organization is sufficient for experience. The framework does not reject functionalism; it specifies that the relevant organization may need to include integrated, temporally persistent, self-regulating, and potentially embodied processes rather than input-output equivalence alone. Whether these features are necessary or sufficient remains an open philosophical and empirical question [13, 14, 18, 36].

8.2 Objection 2: Illusionism

Illusionism poses a deeper challenge by denying that conscious experience exists as ordinarily conceived. If reports of phenomenal qualities can be explained entirely in terms of cognitive representations and dispositions, there may be no genuine valence phenomenon whose organizational basis requires explanation. [48] The framework therefore proceeds conditionally: if conscious experience and valence are genuine, identifying their organizational basis remains a legitimate scientific and ethical objective; if illusionism is correct, the framework loses its target rather than being internally inconsistent. Because welfare and suffering remain central to debates about machine welfare, the framework retains practical relevance while this metaphysical dispute remains unresolved [4, 11, 48].

Conclusion

Artificial valence is best approached through organizational criteria rather than behavioural imitation. The paper proposes four conditions—internal integration, stable but changing internal states, valence–action coupling, and substrate-sensitive implementation, with embodiment and internal regulation as cross-cutting features. These conditions do not establish consciousness or valence; rather, they identify organizational features that can make artificial systems stronger candidates for scientific and ethical investigation.

The framework also supports a precautionary approach to governance. If future systems provide converging evidence of welfare-relevant organization, existing AI governance mechanisms should be capable of incorporating proportionate assessment through system disclosure, independent review, monitoring, and welfare-risk evaluation. The paper's contribution is therefore a provisional interdisciplinary framework for investigating artificial valence and informing responsible governance under continuing uncertainty.

Declarations

Author Contribution

L.I is the sole author of the manuscript.

References

1. Russell, S.J.: Human compatible: artificial intelligence and the problem of control. Viking, New York (2019)
2. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. (2021). arXiv:2108.07258
3. Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., et al.: Consciousness in artificial intelligence: insights from the science of consciousness. arXiv:2308.08708. (2023). https://doi.org/10.48550/arXiv.2308.08708
4. Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., Chalmers, D.: Taking AI welfare seriously. (2024). arXiv:2411.00986
5. Birch, J.: The edge of sentience: risk and precaution in humans, other animals, and AI. Oxford University Press, Oxford (2024). https://doi.org/10.1093/9780191966729.001.0001
6. OECD: What are the OECD principles on AI? OECD Observer. (2019). https://doi.org/10.1787/6ff2a1c4-en
7. UNESCO: Recommendation on the ethics of artificial intelligence. UNESCO, Paris (2021). https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence
8. Tabassi, E.: Artificial intelligence risk management framework (AI RMF 1.0). NIST AI 100-1. National Institute of Standards and Technology, Gaithersburg, MD. (2023). https://doi.org/10.6028/NIST.AI.100-1
9. European Parliament and Council of the European Union: Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union L, 12 July 2024. ELI: (2024). http://data.europa.eu/eli/reg/2024/1689/oj
10. Metzinger, T.: Artificial suffering: an argument for a global moratorium on synthetic phenomenology. J. Artif. Intell. Conscious. **8**(1), 43–66 (2021). https://doi.org/10.1142/S270507852150003X
11. Sebo, J., Long, R.: Moral consideration for AI systems by 2030. AI Ethics. **5**, 591–606 (2023). https://doi.org/10.1007/s43681-023-00379-1
12. Nagel, T.: What is it like to be a bat? Philosophical Rev. **83**(4), 435–450 (1974). https://doi.org/10.2307/2183914
13. Block, N.: Consciousness, accessibility, and the mesh between psychology and neuroscience. Behav. Brain Sci. **30**(5–6), 481–499 (2007). https://doi.org/10.1017/S0140525X07002786
14. Chalmers, D.J.: The conscious mind: in search of a fundamental theory. Oxford University Press, Oxford (1996)
15. Chalmers, D.J.: Could a large language model be conscious? Boston Review, 9 August 2023. (2023). https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/
16. Putnam, H.: Philosophy and our mental life. In: Mind, language and reality: philosophical papers, vol. 2, pp. 291–303. Cambridge University Press, Cambridge (1975).

https://doi.org/10.1017/CBO9780511625251.016

17. Fodor, J.A.: Special sciences (or: the disunity of science as a working hypothesis). Synthese. **28**(2), 97–115 (1974). https://doi.org/10.1007/BF00485230
18. Searle, J.R.: Minds, brains, and programs. Behav. Brain Sci. **3**(3), 417–457 (1980). https://doi.org/10.1017/S0140525X00005756
19. Searle, J.R.: The rediscovery of the mind. MIT Press, Cambridge, MA (1992)
20. Schneider, S.: Artificial you: AI and the future of your mind. Princeton University Press, Princeton (2019). https://doi.org/10.1515/9780691197777
21. Ladak, A.: What would qualify an artificial intelligence for moral standing? AI Ethics. **4**, 213–228 (2024). https://doi.org/10.1007/s43681-023-00260-1
22. Mullally, N.: The self-preservation test for artificial sentience. AI Ethics. **6**, 142 (2026). https://doi.org/10.1007/s43681-026-00983-x
23. Wolfson, I.: Informed consent for AI consciousness research: a Talmudic framework for graduated protections. AI Ethics. **6**, 20 (2026). https://doi.org/10.1007/s43681-025-00852-z
24. Wilks, M., Ladak, A., Loughnan, S.: Why AI might not gain moral standing: lessons from animal ethics. AI Ethics. **6**, 74 (2026). https://doi.org/10.1007/s43681-025-00919-x
25. Tait, I.: Ethically enslaving AI. AI Ethics. **6**, 183 (2026). https://doi.org/10.1007/s43681-026-01049-8
26. Placani, A.: Anthropomorphism in AI: hype and fallacy. AI Ethics. **4**, 691–698 (2024). https://doi.org/10.1007/s43681-024-00419-4
27. Tye, M.: Ten problems of consciousness: a representational theory of the phenomenal mind. MIT Press, Cambridge, MA (1995)
28. Block, N.: On a confusion about a function of consciousness. Behav. Brain Sci. **18**(2), 227–247 (1995). https://doi.org/10.1017/S0140525X00038188
29. Weiskrantz, L.: Blindsight. Clarendon, Oxford (1986)
30. Laureys, S., Pellas, F., Van Eeckhout, P., Ghorbel, S., Schnakers, C., Perrin, F., et al.: The locked-in syndrome: what is it like to be conscious but paralyzed and voiceless? Prog. Brain Res. **150**, 495–511 (2005). https://doi.org/10.1016/S0079-6123(05)50034-7
31. Sutton, R.S., Barto, A.G.: Reinforcement learning: an introduction, 2nd edn. MIT Press, Cambridge, MA (2018)
32. Robinson, M.J.F., Berridge, K.C.: Instant transformation of learned repulsion into motivational wanting. Curr. Biol. **23**(4), 282–289 (2013). https://doi.org/10.1016/j.cub.2013.01.016
33. LeDoux, J.: Rethinking the emotional brain. Neuron. **73**(4), 653–676 (2012). https://doi.org/10.1016/j.neuron.2012.02.004
34. Floridi, L., Cowls, J.: Ethics of artificial intelligence. In: Zalta, E.N., Nodelman, U. (eds.) The Stanford encyclopedia of philosophy, Summer 2022 edn. Stanford University (2022)
35. Dignum, V.: Responsible artificial intelligence: how to develop and use AI in a responsible way. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30371-6

36. Block, N.: Troubles with functionalism. Minnesota Studies in the Philosophy of Science 9:261–325. University of Minnesota Press, Minneapolis (1978)
37. Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A.M., Marshall, W., et al.: Integrated information theory (IIT) 4.0: formulating the properties of phenomenal existence in physical terms. PLoS Comput. Biol. **19**(10), e1011465 (2023). https://doi.org/10.1371/journal.pcbi.1011465
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems 30, pp. 5998–6008. Curran Associates, Inc., Red Hook, NY (2017)
39. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in neural information processing systems 35, pp. 27730–27744. Curran Associates, Inc., Red Hook, NY (2022). https://doi.org/10.52202/068431-2011
40. Seth, A.K., Tsakiris, M.: Being a beast machine: the somatic basis of selfhood. Trends Cogn. Sci. **22**(11), 969–981 (2018). https://doi.org/10.1016/j.tics.2018.08.008
41. Indiveri, G., Liu, S.-C.: Memory and information processing in neuromorphic systems. Proceedings of the IEEE 103(8):1379–1397. (2015). https://doi.org/10.1109/JPROC.2015.2444094
42. Friston, K.: The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. **11**(2), 127–138 (2010). https://doi.org/10.1038/nrn2787
43. Baars, B.J.: A cognitive theory of consciousness. Cambridge University Press, Cambridge (1988)
44. Dehaene, S.: Consciousness and the brain: deciphering how the brain codes our thoughts. Viking, New York (2014)
45. Rosenthal, D.M.: Consciousness and mind. Oxford University Press, Oxford (2005). https://doi.org/10.1093/oso/9780198236979.001.0001
46. Clark, A.: Surfing uncertainty: prediction, action, and the embodied mind. Oxford University Press, Oxford (2016)
47. Barrett, A.B., Mediano, P.A.M.: The phi measure of integrated information is not well defined for general physical systems. J. Conscious. Stud. **26**(1–2), 11–20 (2019)
48. Frankish, K.: Illusionism as a theory of consciousness. J. Conscious. Stud. **23**(11–12), 11–39 (2016)
49. Putnam, H.: Psychological predicates. In: Capitan, W.H., Merrill, D.D. (eds.) Art, mind, and religion, pp. 37–48. University of Pittsburgh, Pittsburgh (1967)
50. Fodor, J.A.: Psychological explanation: an introduction to the philosophy of psychology. Random House, New York (1968)
51. Lau, H., Rosenthal, D.: Empirical support for higher-order theories of conscious awareness. Trends Cogn. Sci. **15**(8), 365–373 (2011). https://doi.org/10.1016/j.tics.2011.05.009

52. Varela, F.J., Thompson, E., Rosch, E.: The embodied mind: cognitive science and human experience. MIT Press, Cambridge, MA (1991)
53. Noë, A.: Action in perception. MIT Press, Cambridge, MA (2004)