Now You See It, Now You Don't: A Theory of Machine Consciousness Must Explain Illusions

Gabriel Simmons
University of California, Davis
gsimmons@ucdavis.edu

Abstract

As machine intelligences become more sophisticated, individuals and societies will encounter myriad attributions of consciousness to machines. Some of these attributions may be correct; many will be illusory. This paper argues that computationalist functionalist theories of machine consciousness can and should seek to explain illusions of machine consciousness. A theory that accounts for illusions reduces the risk of producing false positives—theories that explain merely the appearance of consciousness rather than genuine consciousness. Further, given the consequential nature of consciousness attributions, whether veridical or illusory, such a theory helps society navigate issues of blame and recourse when mistaken attributions cause harm.

Introduction

Some humans find it plausible that today's AI systems are conscious (Guingrich and Graziano 2025; Colombatto and Fleming 2024). Top AI labs are investigating the possibility (Anthropic 2024). Humans now form intimate relationships with AIs (Chu et al. 2025; Laestadius et al. 2024). This would all be appropriate if today's AIs were genuinely conscious. Rather than accepting this radical conclusion, alternative philosophical positions account for the appearance of consciousness in today's AIs without granting genuine consciousness. Technology company executives (Knight 2025) as well as philosophers (Seth 2026) argue that apparent consciousness in today's AIs is merely an *illusion*. There are good reasons to be attracted to this illusionist stance. Illusionism about machine consciousness avoids the moral, social, ethical, and economic consequences of genuine consciousness attribution. Because illusions function socially as face-saving explanations for misperceptions, the illusionist stance also exculpates those who perceive consciousness in machines from an otherwise blameworthy error. Illusory or not, consciousness attributions are highly consequential. An entity's consciousness status is intimately linked with its status as a living being, as a moral patient, and as a holder of individual rights. We consider illusions of machine consciousness from a philosophical perspective, through the lens of computationalist functionalism (Bach and Sorenson 2025; Piccinini 2010), arguing that illusions can and should be first-class objects of inquiry for a computationalist functionalist research program.

Computationalist Functionalism

Computational systems are systems that can be adequately described in terms of state transitions (Bach and Sorenson 2025; Piccinini 2010). Functionalism argues that the consciousness of a system depends on its functional organization (Bach and Sorenson 2025), rejecting the essentialist idea that consciousness is substrate-dependent. Computationalist functionalism (CF) about consciousness says roughly that consciousness is the name for an equivalence class of patterns of information processing (Bach and Sorenson 2025; Piccinini 2010).

Illusions and Illusionism

Philosophically, illusionism about P is the position that P seems to be true but is not actually true. Strong illusionism about consciousness is the position that consciousness *seems* to exist, but does not in fact exist (Frankish 2016). Weak illusionist views are illusionist positions about particular aspects of consciousness rather than consciousness in its entirety. A weak illusionist might argue, for example, that there really are phenomenal properties, but they are not actually ineffable (Frankish 2016).

Socially, not all misperceptions are commonly regarded as illusions. We often say we are affected by illusions when we are affected by misperceptions that are common to the kind of minds we have. This grants illusions an exculpatory social function. Recognition of a phenomenon as an illusion rather than merely a misperception shifts blame from the observer to the phenomenon itself, or to the observer's membership in a broader class of minds. An illusion is a ternary relation, a tendency of a certain class of observers to experience a certain kind of misperception of a certain kind of phenomena.

Illusionist positions posit a misperception and an explanation for that misperception. The misperception component says something like "our attributions of consciousness to ourselves or others are not veridical". The explanatory component says something like "illusion status explains why our attributions of consciousness to ourselves or others are widespread, despite their non-veridicality".

Illusionism and Machine Consciousness

The stance that yields illusionism about human consciousness can also be applied to machines, producing various forms of illusionism about machine consciousness. *Strong illusionism about machine consciousness* is the view that machine introspection seems to be phenomenally conscious but is not actually so. *Weak illusionism about machine consciousness* says that machine consciousness exists, but some of its properties are illusory. We may also consider illusionist positions about particular classes of machines, e.g. "while machine consciousness may be real, apparent consciousness in large language models is an illusion".

Computationalist functionalism about consciousness holds that consciousness is a pattern of functional organization that can be formalized computationally. Of course, one can hold a computationalist functionalist view towards other cognitive tendencies, including illusions.

At first glance, one might think that illusionism and computationalist functionalism are mutually exclusive in a simple sense. All CF needs to do is adopt a metaphysical stance that assumes that consciousness is real, paint illusionism as the view that consciousness is not real, and be done, agreeing to disagree at the level of broad-strokes metaphysics. The interface between the two positions may be more granular than that; the following section sketches why this might be.

A Theory of Machine Consciousness Should Explain Illusions

There are two scopes in which a research program about machine consciousness can contend with illusions. The first scope is internal, focused on the research program's own products. Here, the argument for attention to illusions is philosophical. It is logically possible that there are real patterns of functional organization that can be formalized computationally and yet only appear to be phenomenal without actually being so. Such patterns might elicit widespread attributions of phenomenality from observers when they are implemented in machines, and yet still not have phenomenal properties. There may be theories that claim to have identified computational patterns that underlie consciousness, and yet have only identified computational patterns that underlie merely apparent consciousness. To claim to have succeeded in explaining or producing real consciousness, a research program must claim that its products have real phenomenal properties, and not merely appearances of phenomenal properties. Alternatively, the program may settle for having produced an explanation for apparent phenomenal properties that are only possibly real. In this sense, the research program either settles for explaining the possibly-phenomenal, or contends with false positive risk. In other words, proponents of any particular computationalist functionalist theory of machine consciousness will have to contend with skeptics who say that the theory explains only illusions of machine consciousness; to overcome the skeptic, the research program must explain why its products are not illusions.

The second scope is external, focused on the public's judgments of machine consciousness. Here, the argument for attention to illusions is social and ethical. In a nutshell, attributions of consciousness to machines are likely to be widespread and highly consequential, and it is likely that some of these will be illusory. Some philosophers argue that this already occurs (Seth 2026; Dennett 2023; Frankish 2024). Furthermore, many individuals and groups will allege that machine consciousness attributions are the result of illusion, regardless of their actual status as veridical or illusory. There are social and economic pressures encouraging illusionist stances; the discovery of machine consciousness would be a significant disruption to the status quo, likely resulting in regulatory, social, or other costs to the major producers of machine intelligences. Illusionism explains away these otherwise disruptive phenomena.

One might ask why computationalist functionalism should concern itself with this second scope. One might argue that the normatively significant questions about illusory consciousness attribution are better addressed by other fields like cognitive psychology, social epistemology, or HCI, as one reviewer suggests. This may be the case if the boundary between illusory attribution and veridical attribution is clear. Cognitive psychologists and social epistemologists may certainly offer explanations for why people tend to attribute consciousness to certain kinds of machines, and HCI researchers contribute to our understanding and control of the kind of interfaces that elicit such attributions. But whether these attributions are veridical or illusory is a question that demands a theory of machine consciousness. Among many stances towards consciousness, computationalist functionalism is particularly well-suited to practical application (monitoring, regulation, adjudication), since its theories of consciousness are grounded in computational patterns that can be formalized and implemented in machines.

Conclusion

As machine intelligences become more sophisticated, individuals and societies will encounter myriad illusions of machine consciousness, perhaps well before genuine machine consciousness is achieved. Some already attribute consciousness to machine systems, while others already allege that these attributions are the result of illusion. Computationalist functionalism says that consciousness is a functionally-defined, substrate-independent pattern that can be formalized computationally. By incorporating explanations for illusions into its theories, a computationalist functionalist research program strengthens its output against skeptical counterargument and false positive risk. Further, many of the societal consequences of attributed machine consciousness will stem from the attribution itself (veridical or illusory), and many others will stem from the determination between the two. A computationalist functionalist research program is uniquely equipped to draw a precisely testable boundary around genuinely conscious machines, and another around those that are merely apparently conscious, serving as a foundation for monitoring, regulation, and adjudication.

Acknowledgements

The author thanks Oisín Hugh Clancy and the anonymous reviewers for input that contributed to the ideas in this work.

References

Anthropic. 2024. Exploring Model Welfare. https://www.anthropic.com/research/exploring-model-welfare.

Bach, J.; and Sorenson, H. 2025. The Machine Consciousness Hypothesis. Technical report, California Institute for Machine Consciousness.

Chu, M. D. H.; Gerard, P.; Pawar, K.; Bickham, C.; and Lerman, K. 2025. Illusions of Intimacy: Emotional Attachment and Emerging Psychological Risks in Human-AI Relationships. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Colombatto, C.; and Fleming, S. M. 2024. Folk Psychological Attributions of Consciousness to Large Language Models. *Neuroscience of Consciousness*, 2024(1): niae013.

Dennett, D. C. 2023. The Problem With Counterfeit People. *The Atlantic*.

Frankish, K. 2016. Illusionism as a Theory of Consciousness. *Journal of Consciousness Studies*, 23(11-12): 11–39.

Frankish, K. 2024. What Are Large Language Models Doing? In Strasser, A., ed., *Anna's AI Anthology: How to Live with Smart Machines?*, 55–78. Berlin: Xenomoi.

Guingrich, R. E.; and Graziano, M. S. A. 2025. Chatbots as Social Companions: How People Perceive Consciousness, Human Likeness, and Social Health Benefits in Machines. *Oxford Intersections: AI in Society*.

Knight, W. 2025. Microsoft's AI Chief Says Machine Consciousness Is an 'Illusion'. *Wired*.

Laestadius, L.; Bishop, A.; Gonzalez, M.; Illenčík, D.; and Campos-Castillo, C. 2024. Too Human and Not Human Enough: A Grounded Theory Analysis of Mental Health Harms from Emotional Dependence on the Social Chatbot Replika. *New Media & Society*, 26(10): 5923–5941.

Piccinini, G. 2010. The Mind as Neural Software? Understanding Functionalism, Computationalism, and Computational Functionalism. *Philosophy and Phenomenological Research*, 81(2): 269–311.

Seth, A. 2026. The Mythology Of Conscious AI. *Noema Magazine*.