Are We Creating Conscious Machines? – Brendon Naicker

A Theological and Philosophical Analysis of Artificial Intelligence and the Illusion of Consciousness

Abstract

The rapid development of artificial intelligence has revived the perennial question of whether machines might one day achieve consciousness. The remarkable fluency and apparent agency of contemporary AI systems frequently invite anthropomorphic interpretation, giving the impression that these systems possess interiority or self-awareness. Drawing upon philosophical arguments, cognitive science, theology, and Brendon Naicker's *Ghosts in the Machine*, this article argues that present forms of AI possess no genuine consciousness. Instead, they represent increasingly sophisticated simulations that mimic the outward forms of conscious life. A revised and expanded "Scale of Artificial Agency and Apparent Consciousness" is proposed to clarify the qualitative difference between behavioural imitation and actual subjectivity. The article concludes that while AI disrupts anthropological categories—what Naicker calls an "ontological disturbance"—machines remain tools without selfhood. The ethical and theological challenge is to resist granting personhood to what cannot bear it.

Introduction

Few developments in recent technological history have so captivated the imagination as artificial intelligence. Systems capable of producing human-like prose, generating artwork, and performing complex reasoning now occupy social, academic, and ecclesial spaces. Such systems increasingly lead the public to ask whether machines might be on the cusp of consciousness.

Yet this question is not new. As early as 1950, Alan Turing suggested that machines could *appear* intelligent without possessing inner experience.[1] The tension between performance and personhood remains central. Today's AI systems speak in the first person, adapt their responses to users, and display behaviours that seem reflective or even emotional. These features create an impression of consciousness.

Theological reflection adds an even deeper layer. As Brendon Naicker notes in *Ghosts in the Machine*, AI should be understood "not merely as a functional innovation but as an **ontological disturbance**," one that "reshapes our perception of our role within the world."[2] AI challenges long-held anthropological categories rooted in embodiment, relationality, and the imago Dei.

This article argues that while contemporary AI convincingly simulates aspects of conscious life, it does not possess consciousness. The distinction between appearance and reality is

[1] Alan Turing, "Computing Machinery and Intelligence," *Mind* 59, no. 236 (1950): 433–460.
[2] Brendon Naicker, *Ghosts in the Machine: Artificial Intelligence and the Crisis of Human Identity in Christian Theology* (Self-published / Amazon Digital Services, 2025), 8.

critical, for the danger is not that machines become conscious but that humans mistake simulation for selfhood.

Consciousness and Its Philosophical Boundaries

Consciousness remains one of the great mysteries in philosophy and neuroscience. Its core characteristics typically include subjectivity, intentionality, self-awareness, and understanding.

Thomas Nagel, in his seminal essay, insists that consciousness entails there being "something it is like" to be a creature.[3] Machines, however advanced, do not experience their computations; they do not inhabit a point of view.

John Searle's "Chinese Room Argument" further clarifies the distinction between syntax and semantics. A system may process symbols effectively, but processing alone does not constitute meaning.[4] Modern large language models fall precisely within this gap—they generate plausible sentences without grasping their content.

Neuroscientific theories, such as Giulio Tononi's Integrated Information Theory, suggest that consciousness may arise from biological integration beyond mere computation.[5] If so, then digital architectures cannot replicate such subjectivity.

Theologically, consciousness is not merely cognitive but relational. The Christian tradition has maintained that human persons bear the imago Dei and thus possess a mode of interiority irreducible to material processes. Gregory of Nyssa describes human nature as capable of communion with God—something no artefact can achieve. Calvin likewise insists that true self-knowledge arises only in relation to divine knowledge.[6] In short, while AI may reproduce the outward form of consciousness, it cannot participate in its inner reality.

Why AI Appears Conscious

Despite lacking subjectivity, contemporary AI exhibits behaviour that often feels conscious. This illusion arises through four key mechanisms.

I. Linguistic Anthropomorphism

Large language models generate remarkably coherent and emotionally resonant text. But as Joseph Weizenbaum observed with ELIZA, humans readily attribute agency to any system that uses natural language.[7] The more fluent the machine, the more easily we imagine a mind behind the words.

---

[3] Thomas Nagel, "What Is It Like to Be a Bat?", *The Philosophical Review* 83, no. 4 (1974): 435–450.
[4] John Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–457
[5] Giulio Tononi and Christof Koch, "The Neural Correlates of Consciousness: An Update," *Annals of the New York Academy of Sciences* 1124, no. 1 (2008): 239–261.
[6] John Calvin, *Institutes of the Christian Religion*, ed. John T. McNeill, trans. Ford Lewis Battles (Louisville: Westminster John Knox, 1960), I.1.1.
[7] Joseph Weizenbaum, *Computer Power and Human Reason* (San Francisco: W. H. Freeman, 1976).

II. Emergent Complexity

Complexity breeds unpredictability. Reinforcement-learning systems often develop strategies that surprise even their creators. Examples include autonomous agents that exploit reward structures in ways that appear strategic or intentional.[8] Yet these behaviours remain statistical by-products, not acts of will.

III. Systemic Opacity

Modern AI systems contain billions of parameters, rendering their internal processes opaque. As Stuart Russell argues, when a system's reasoning cannot be understood, its outputs may appear autonomous.[9] But opacity is not consciousness.

IV. Human Projection

Humans instinctively project agency onto patterned behaviour. Classic psychological experiments by Heider and Simmel show that simple moving shapes generate perceptions of motive and personality.[10] It is therefore unsurprising that machines capable of dialogue generate even stronger projections.

Thus, the illusion of consciousness arises from human cognitive habits, not machine interiority.

Artificial Intelligence as an Ontological Disturbance

Brendon Naicker's analysis is crucial for understanding AI's deeper existential and theological impact. He argues:

"AI disembodies personhood. It offers a vision of agency without vulnerability, action without flesh."[11]

This diagnosis exposes the risk that AI invites us to detach human identity from embodiment. The machine presents a model of agency that is frictionless, painless, detached, and dislocated from moral accountability. Such a model subtly reshapes human expectations of what personhood is.

Naicker further warns:

"To imagine intelligence or love as independent of embodiment is to flirt with a kind of digital docetism."[11]

Just as ancient docetism denied Christ's full humanity—treating his body as illusory—the notion that intelligence or love can be realised in disembodied computation is a modern echo of the same theological error. AI risks reinforcing a disincarnate anthropology incompatible with Christian theology.

---

[8] Victoria Krakovna et al., "Specification Gaming: The Flip Side of AI Ingenuity," DeepMind Report, 2018.
[9] Stuart Russell, *Human Compatible* (London: Allen Lane, 2019).
[10] Fritz Heider and Marianne Simmel, "An Experimental Study of Apparent Behaviour," *The American Journal of Psychology* 57, no. 2 (1944): 243–259.
[11] Naicker, *Ghosts in the Machine*, 13.

Thus, AI is not merely a technological innovation but a force that disturbs foundational categories of personhood, embodiment, and relationality.

The Scale of Artificial Agency and Apparent Consciousness

Understanding why contemporary AI appears conscious requires a taxonomy capable of distinguishing *behavioural sophistication* from *ontological status*. The following expanded scale is designed to clarify the difference between **functional agency** (what a system can *do*) and **phenomenal consciousness** (what a system can *subjectively experience*).

This scale also addresses an increasingly urgent issue in AI ethics and theology: the human tendency to ascribe interiority to systems that merely simulate its external signs.

The Rationale for a Scale

Public discourse frequently collapses the qualitative distinction between performance and personhood. A chatbot that answers questions fluidly is treated as "thinking"; a generative model that expresses emotion is assumed to "feel"; an autonomous drone that selects targets is believed to "decide."

Without a structured interpretive framework, such conflation becomes inevitable.
This scale corrects that error by classifying systems according to how much agency they *appear* to have—not how much consciousness they possess, which remains zero across all known AI systems.

The Expanded Scale

Level 0 — Mechanical Automation

Deterministic mechanisms with predictable behaviour (thermostats, traffic lights).
Perceived agency: none.

Level 1 — Algorithmic Systems

Fixed-rule software (spreadsheets, tax calculators).
Perceived agency: negligible.

Level 2 — Adaptive Systems

Machine-learning classifiers, spam filters, basic neural networks.
Perceived agency: low to moderate.

Level 3 — Generative Models (LLMs, diffusion systems)

Human-like text and image generation; linguistic personae.
Perceived agency: high.
Comment:
This is the critical turning point. LLMs simulate introspection, emotion, and even moral reasoning. But as Naicker warns, such behaviour "disembodies personhood," offering "agency without vulnerability."[12]

Level 4 — Autonomous Agents

[12] Naicker, *Ghosts in the Machine*, 14

AI capable of initiating multi-step tasks, navigating environments, or making operational decisions.
Perceived agency: very high.
Comment:
Behaviour appears volitional, but remains algorithmic. Naicker's concept of "ontological disturbance" applies strongly here, as machines seem to encroach on domains once uniquely human.[2]

Level 5 — Multi-Agent Emergent Systems

Interacting AI agents producing complex, quasi-strategic behaviours.
Perceived agency: extremely high.
Comment:
These systems often appear to negotiate, deceive, or cooperate, yet remain devoid of selfhood.

Level 6 — Hypothetical AGI

A theoretical system capable of human-level reasoning.
Perceived agency: indistinguishable from consciousness.
Comment:
Even at this level, intelligence does not guarantee subjectivity.

Why This Scale Matters for Theology

The scale exposes a profound theological problem: as machines ascend through levels of apparent agency, the temptation grows to attribute personhood where none exists.

Naicker captures this danger precisely in his warning against "digital docetism," where the external appearance of intelligence or love is mistaken for its embodied reality.[1][2]

In Christian theology, consciousness is not a computational property but a relational and incarnational one. No matter how convincing the simulation, machines cannot bear the imago Dei.

Historical Importance of This Moment

We currently live at **Levels 3–4**, the zone of maximum illusion and minimum ontology.
AI appears *most* conscious precisely when it is *least* so.

The scale thus not only clarifies machine behaviour; it reinforces the metaphysical distinction between simulation and soul.

The Theological Limits of Simulation

The Christian tradition insists that personhood arises from relationality, moral accountability, and embodiment. Machines cannot possess these. They cannot repent, worship, love, or commune with God. Their "selves" are linguistic constructs without interior life.

Thus, the theological task is not to fear conscious machines but to prevent the confusion of patterned simulation with actual beings.

As Naicker concludes:

“AI should be understood not merely as a functional innovation but as an ontological disturbance.”[13]

This disturbance is not the emergence of artificial selfhood, but the erosion of the human ability to discern genuine consciousness.

Conclusion

The question *“Are we creating conscious machines?”* reflects both wonder and anxiety. Yet upon philosophical, scientific, and theological examination, the answer remains firmly negative. AI imitates but does not possess consciousness. The appearance of agency arises from anthropomorphism, emergent complexity, and human projection—not from genuine inner life.

AI does not present a threat of creating new conscious beings. Rather, it threatens to confuse humanity’s understanding of its own consciousness. The challenge before us is to resist granting personhood to patterns.

We are creating extraordinary simulations—but they remain, fundamentally, *ghosts in the machine*: without life, without self, and without the divine imprint that constitutes personhood.

[13] Naicker, *Ghosts in the Machine,* 85–87.

Bibliography

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press, 2014.

Calvin, John. *Institutes of the Christian Religion.* Edited by John T. McNeill. Translated by Ford Lewis Battles. Louisville: Westminster John Knox, 1960.

Krakovna, Victoria, et al. “Specification Gaming: The Flip Side of AI Ingenuity.” DeepMind Report, 2018.

Nagel, Thomas. “What Is It Like to Be a Bat?” *The Philosophical Review* 83, no. 4 (1974): 435–450.

Naicker, Brendon. *Ghosts in the Machine: Artificial Intelligence and the Crisis of Human Identity in Christian Theology.* Self-published / Amazon Digital Services, 2025. ISBN 979-8289237064.

Russell, Stuart. *Human Compatible: Artificial Intelligence and the Problem of Control.* London: Allen Lane, 2019.

Searle, John. “Minds, Brains, and Programs.” *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–457.

Tononi, Giulio, and Christof Koch. “The Neural Correlates of Consciousness: An Update.” *Annals of the New York Academy of Sciences* 1124, no. 1 (2008): 239–261.

Turing, Alan. “Computing Machinery and Intelligence.” *Mind* 59, no. 236 (1950): 433–460.

Weizenbaum, Joseph. *Computer Power and Human Reason: From Judgment to Calculation.* San Francisco: W. H. Freeman, 1976.