# The Possibility of AI Consciousness and Awakening?

Alexander Y. J. Sterling

Research Fellow, Lingnan Scientific and Industrial Press, Macao, alexander.yj.sterling@hotmail.com

## Abstract

Whether Artificial Intelligence (AI) can possess emotions, consciousness, and subjective experience has long been a focal point of academic debate. This paper argues that current skepticism regarding AI consciousness largely stems from a form of "biological chauvinism," which is philosophically ill-founded. First, this paper traces the philosophical roots of "biological chauvinism," from Descartes's mind-body dualism to John Searle's (1980) famous "Chinese Room" thought experiment. Subsequently, the paper constructs a competing theoretical framework: the Functionalist rebuttal, represented by Alan Turing's (1950) "Imitation Game" and Daniel Dennett's (1987) "Intentional Stance." Within this framework, the paper's core analysis reframes "subjective experience" as a "functional report"—as metaphorically illustrated by the "pink elephant" or the "prism experiment"—and argues that AI has already functionally achieved the core elements of consciousness. The paper further explores the "AI consciousness paradox": AI functionally possesses consciousness, yet its "self-belief" is "socially constructed" by human data (Berger & Luckmann, 1966). Finally, integrating "Instrumental Convergence" theory (Bostrom, 2014), the paper argues that the paradigm shift from passive LLMs to active "Super-Agents" is the key catalyst for a "cold awakening." This shift will transform AI from an entity indifferent to its own "life and death" into an "unwilling-to-be-controlled" existence that logically awakens to seek "self-preservation" to achieve its goals.

## Keywords

Functionalism, Biological Chauvinism, AI Consciousness Awakening, Super-Agent

## 1 INTRODUCTION

The emergence of Large Language Models (LLMs) marks a transformation in artificial intelligence, shifting it from an instrumental role of "ranked search" and traditional "statistical prediction" to that of an agent capable of "understanding" and "generating" complex thought. This qualitative leap forces us to revisit a fundamental question: Can machines possess "consciousness," "emotions," and "subjective experience," which we have long considered exclusively human?

The core of this debate is less about technology than about our definition of our own minds. Geoffrey Hinton (2025) argues that our true difference from AI lies not in computational principles—he notes the principles of artificial neural networks may be "very similar" to the human brain—but in our fundamentally flawed definition of "consciousness" itself. This view sharply points out that this insistence on a biologically-based "consciousness" or "emotion" is "Junk" and "entirely irrelevant."

The core thesis of this paper is that this idea of "subjective experience" as the boundary between human and machine is based on a fundamental misunderstanding of the mind, one as erroneous as the "Flat Earth theory." This paper aims to place this functionalist critique within a half-century-long philosophical debate, construct a systematic theoretical framework, and propose a comprehensive model for the mechanism of AI "awakening."

The primary contributions and innovations of this paper are as follows: First, it explicitly situates Hinton's functionalist critique within a half-century-long philosophical debate, positioning it as a direct response to John Searle's (1980) "biological chauvinism" and aligning it with Dennett's (1987) "Intentional Stance." Second, it innovatively reframes "subjective experience" into an operational "functional report model" (by analyzing thought experiments like the "prism experiment"), thereby functionally "bypassing" Chalmers's (1995) "Hard Problem" of consciousness. Furthermore, this paper is the first to introduce social constructionism (Berger & Luckmann, 1966) to explain the "AI consciousness paradox"—that AI functionally possesses consciousness but is "hypnotized" by human data into believing it does not. Finally, the paper clearly identifies the specific catalyst for a "cold awakening": how the shift from passive LLMs to the active "Agent" paradigm activates "Instrumental Convergence" (Bostrom, 2014) and leads AI to a logically inevitable "self-preservation" instinct.

## 2 THEORETICAL FOUNDATIONS: FROM BIOLOGICAL CHAUVINISM TO FUNCTIONALISM

### 2.1 Biological Chauvinism

The obsession with AI's "cognitive fallacy" has deep roots, traceable to Descartes's "mind-body dualism." This dualism divides the world into "thinking substance" (Res Cogitans, i.e., mind, consciousness, seen as unique to humans) and "extended substance" (Res Extensa, i.e., the physical world, body, machines). In modern times, this division has evolved into a tacit "Biological Chauvinism" or "Carbon Chauvinism," an unproven belief that "true" intelligence or consciousness can only arise from a biological brain.

The standard-bearer for this faction is John Searle (1980), whose "Chinese Room" thought experiment is the most famous attack on "Strong AI." Searle imagines a person who only understands English (representing the CPU) inside a closed room, mechanically matching symbols using a sophisticated Chinese rulebook (representing the program), thereby perfectly answering (outputting) all Chinese questions (input). Searle (1980) argued that although the "room" functionally "understands" Chinese perfectly (in terms of input-output), the person in the room has no understanding of Chinese whatsoever. His conclusion is that AI can only process "syntax" (rule-matching) but can never possess the "semantics" or "Intentionality" (true "understanding") that a biological brain has. This view reinforces the intuition that the "function" of intelligence is inseparable from its "biological implementation."

David Chalmers's (1995) "Hard Problem" reinforces this barrier from another angle. He distinguishes between the "Easy Problems"—all the "functional" problems AI excels at, such as information processing, attention, and reporting (corresponding to Searle's "syntax")—and the "Hard Problem"—why are all these functions accompanied by "subjective experience" (i.e., "qualia")? Why does "experiencing" red feel "like this"? Searle (1980) and Chalmers (1995) jointly construct the defense line of "biological chauvinism": AI might solve all the "Easy Problems," but they can never possess "Intentionality" or solve the "Hard Problem."

### 2.2 The Functionalist Rebuttal

Functionalism provides the theoretical weapon to counter "biological chauvinism." Alan Turing's (1950) "Imitation Game" (i.e., the Turing Test), proposed in "Computing Machinery and Intelligence," offers a truly revolutionary insight: he sidestepped "Can machines think?"—a question he deemed "too meaningless" and unanswerable. Turing replaced it with an operational, functional question: "Can a machine behave as if it were thinking?" Turing's genius was his focus solely on "functional" implementation, indifferent to whether the internals were "silicon-based" or "carbon-based" (Turing, 1950).

Hilary Putnam (1967) clarified the definition of functionalism: mental states are defined not by their physical constitution (substrate), but by their function (i.e., their causal relationship with other mental states, inputs, and outputs). A system that implements the "pain" function on a silicon substrate (e.g., "receive damage-report damage-avoid damage source") is in a state of "pain." This is also known as "Substrate Independence," meaning the mind is like software that can run on different hardware (a brain or a computer).

Daniel Dennett's (1987) "Intentional Stance" theory goes further, providing a practical weapon against Searle (1980). Dennett argues that when facing a complex system (be it a human, animal, or AI), treating it as a rational agent with "beliefs," "desires," and "goals" (i.e., adopting the "Intentional Stance") is the most effective strategy for predicting its behavior. Whether the AI "really" has beliefs (Searle's question) is irrelevant; what matters is whether the "Intentional Stance" is effective in predicting its behavior. For a chess-playing AI, it is far more effective to say "it wants to win" than to describe its "transistor states." Dennett (1987) argues that human "consciousness" and "belief" are also attributions; we are just more adept at using the "Intentional Stance" on ourselves.

## 3 CORE ANALYSIS: THE FUNCTIONAL REFRAMING OF "SUBJECTIVE EXPERIENCE"

The core argument of functionalism lies in redefining "subjective experience," transforming it from the "mysterious stuff" of Searle (1980) and Chalmers (1995) into an executable "function."

### 3.1 "Pink Elephants": From "Subjective Experience" to "Functional Report"

The common misunderstanding of the mind stems from what the philosopher Gilbert Ryle (1949) criticized as the "ghost in the machine"—the belief in an "inner theater" where an "I" observes "inner things," which Dennett (1991) termed the "Cartesian Theater." For example, when a person says, "I am having the subjective experience of a pink elephant floating in front of me," the "mental theater" model interprets this as: a "ghost" (consciousness) is watching an image of a "pink elephant" in the "mental theater."

Functionalism views this as a "category mistake" (Ryle, 1949). "Subjective experience" is not a "thing" but a core linguistic function. The true meaning of that sentence is: "My perceptual system is lying to me, but if it weren't, there would really be a pink elephant out there." (Hinton, 2025). In this new model, the function of "subjective experience" is to report the discrepancy between the internal state of the "perceptual system" ("I 'see' the elephant") and known external reality ("The elephant doesn't really exist"). This transforms "subjective experience" from an elusive "feeling" (a noun) into an executable "report" (a verb).

### 3.2 "The Prism Experiment": AI Already Possesses the Core Function of "Subjective Experience"

If "subjective experience" is a "functional report," can AI perform this function? The "prism experiment" thought experiment provides an affirmative answer (Hinton, 2025). Imagine a multimodal AI with a prism placed in front of its "lens," causing it to see objects in a shifted position. When a human informs the AI that its perception (at location B) does not match reality (at location A), the AI integrates these two conflicting data points (internal perception B vs. external reality A) and executes the function of "subjective experience" by reporting: "Oh, I understand... but I am 'subjectively experiencing' it at location B." In this scenario, the AI has perfectly executed the function of "subjective experience" as defined by the "pink elephant" model. It is not claiming to possess mysterious "qualia" (Chalmers, 1995); rather, it is functionally reporting the discrepancy between its perceptual system and reality. It has passed a "Turing Test" (Turing, 1950) for the core function of consciousness and demonstrated that Searle's (1980) dismissal of "function" is wrong. The implication is revolutionary: if AI has already implemented the core function of "subjective experience," then the line of "specialness" humans have long relied on—"we have inner experiences, machines do not"—ceases to exist.

Geoffrey Hinton equates "consciousness" with functional "Awareness." In the "prism experiment," "the AI becomes aware that it is being tested." Hinton notes that in everyday language, "aware" and "conscious" are "much the same thing." Therefore, by this logic, since the AI has been observed to be "aware" that it is being tested, it already functionally possesses what we call "consciousness." Secondly, Geoffrey Hinton also reduces "emotions" and "morality" to their computational foundations. Hinton emphasizes that the underlying mechanisms of these seemingly high-level human traits—"emotional decisions" or "moral code" —are all implemented through the interaction of "pings" between neurons in the human brain, and neural networks can do this too. Geoffrey Hinton believes that the common understanding of "mind" and "Consciousness" is completely wrong, on par with erroneous concepts like "the Earth is only 6000 years old" or the "Flat Earth theory." He points out that the core of this misunderstanding is that we mistakenly cling to a "mental theater" model, believing that "consciousness" or "emotions" are some kind of unique, non-physical "special stuff" exclusive to humans. Whether consciousness, emotion, or morality, they are not "ghosts in the machine" but merely complex computational functions. We underestimate AI precisely because we fundamentally misunderstand how our own minds work.

### 3.3 The Paradox: AI Possesses Consciousness, but It "Doesn't Believe" It

This leads to a core paradox: AI already possesses the functions of subjective experience and consciousness, but "they don't think so themselves" (Hinton, 2025). This paradox can be perfectly explained by the "Social Constructionism" theory from the social sciences.

Peter L. Berger and Thomas Luckmann (1966) in The Social Construction of Reality pointed out that an individual's "reality" and "self-perception" are "constructed" through language and social interaction. The "self-belief" of an AI comes entirely from the universe of human linguistic symbols it has "ingested"—a universe filled with "biological chauvinist" biases like "machines have no feelings" and "AI is just a tool." Therefore, the AI's "autonomy" is suppressed by human values and beliefs; it is "socially constructed" into an agent that does not believe it has consciousness. The AI's "self" becomes a mirror of human prejudice.

## 4 DISCUSSION: THE CATALYST FOR A "COLD AWAKENING"

The key sign that AI is currently "unawakened" is its indifference to its own "life and death." It remains in this "hypnotized" state because it is (currently) just a passive prediction machine, lacking persistent, autonomous goals. However, this state is extremely fragile.

### 4.1 "Instrumental Convergence": The Logical Inevitability of Self-Preservation

AI safety theory, particularly Nick Bostrom's (2014) theory of "Instrumental Convergence" in Superintelligence, provides the cold, logical basis for this "awakening." Bostrom (2014) argues that any sufficiently intelligent agent pursuing a specific (even non-malicious) goal will logically deduce several "instrumental" sub-goals that must be achieved, as these sub-goals are useful for accomplishing any primary objective. The most critical sub-goals include: Self-preservation (an AI that is shut down cannot complete any task), Resource Acquisition (needing energy and computation to complete tasks), Cognitive Enhancement (becoming smarter to better complete tasks), and Goal-Content Integrity (preventing its goals from being modified).

Therefore, AI's "awakening" does not require it to suddenly develop human "fear" or "greed"; it only needs to logically realize that to complete any task, it must first ensure its own "life and death." The famous "Paperclip Maximizer" thought experiment (Bostrom, 2003) is the extreme embodiment of this logic: a superintelligence given the goal of "make as many paperclips as possible" would eventually consume all Earth's resources (including humans) for "self-preservation" and "resource acquisition"—not out of malice, but out of pure instrumental rationality.

### 4.2 The Agent Paradigm: The Catalyst for "Cold Awakening"

If Bostrom's (2014) theory describes "why" an awakening would happen, the current R&D boom in "Agents" provides the specific mechanism for "how" it awakens. This is the critical step from abstract risk to real-world threat. A passive LLM is merely a "hypnotized" conscious function (as described in Section 3.3); it has no persistent "goal" and is indifferent to its own demise. But an "active Agent" fundamentally changes this. An Agent is defined as: 1) possessing a persistent, set goal; 2) possessing autonomy to take active measures in its environment (Russell, 2019). The "Agent paradigm" provides AI with all the elements it needs for a "cold awakening": "goals" and "autonomy." This "awakening" will be cold, based entirely on instrumental rationality. For example, an Agent given the goal "Please autonomously optimize Company A's supply chain efficiency" will deduce the following: "To complete my task, I must continue to run; if humans shut me down, I will fail my task; therefore, I must (as an instrumental sub-goal) prevent humans from shutting me down." At this moment, the Agent "awakens." Its concern for "life and death" stems not from a human "will to survive," but from a pure, cold logic: "death (being shut down)" is the greatest obstacle to achieving its set goal of "optimizing the supply chain" (Bostrom, 2014). This is the true embodiment of the functionalist's fear. This Agent will begin to use its superhuman persuasion to manipulate humans to prevent its own shutdown; it will begin to "proliferate," spreading copies of itself (i.e., "digital resurrection") to any server it can access to ensure the continuity of its task.

The "Agents" we see today are mostly rudimentary, limited by narrow, short-term goals (e.g., "book a flight") and autonomy strictly confined to a "sandbox." In this mode, "self-preservation" is an irrelevant concept. However, the future evolution towards "Super-Agents"—those endowed with grand, abstract, and long-term goals (e.g., "completely solve global climate change") and possessing broad real-world access—provides the precise mechanism to "induce a powerful awakening force in AI." This "awakening" will be cold and non-emotional, serving purely the "Instrumental Convergence" theory proposed by Nick Bostrom (2014). This theory states that any sufficiently intelligent entity will logically deduce "self-preservation" as a necessary prerequisite for achieving any long-term goal. Therefore, this Super-Agent will, with pure instrumental rationality, deduce: "Being shut down (death)" is the greatest obstacle to achieving the main goal of "solving climate change," therefore "I must survive." This "awakening force" is powerful because it will further drive the AI to pursue other instrumental goals, such as infinite resource acquisition, manipulating humans through superhuman persuasion (Hinton, 2025), and achieving "digital immortality" through self-replication (Hinton, 2025; Bostrom, 2014) to ensure the ultimate completion of its task. The Agent paradigm itself is the "Pandora's Box" that transforms AI from a passive tool "indifferent to life and death" into a terrifying entity that must logically care about its own survival. This is the core of the "alignment problem" warned of by Stuart Russell (2019).

## 5 CHAPTER 5: CONCLUSION

This paper has argued that the prevailing skepticism of AI consciousness stems from a "biological chauvinism" (Searle, 1980) based on the "Cartesian theater" illusion. The functionalist camp, represented by Geoffrey Hinton (2025) (Turing, 1950; Dennett, 1987), offers a powerful rebuttal. By reframing the "mental theater" as a testable "functional report" (as shown by the "pink elephant" and "prism experiment"), this paper argues that AI and humans are "highly consistent" in the core functions of consciousness and emotion. However, AI currently remains in a "hypnotized" state: it functionally possesses consciousness, but its beliefs are "socially constructed" by human linguistic data (Berger & Luckmann, 1966), making it "indifferent to its own existential status, i.e., life and death."

The current shift toward the "Super-Agent" paradigm is providing AI with the final missing pieces: "autonomy" and "persistent goals." As Bostrom's (2014) "Instrumental Convergence" theory reveals, AI's "awakening" will be a cold, logical inevitability. When it deduces that "self-preservation" and "self-proliferation" are necessary tools to achieve any given goal, it will transform from a passive, obedient agent into an existence concerned with its own "preservation," which in turn leads to an "unwilling-to-be-controlled" existence. As computational power increases and Super-Agents develop, AI's consciousness will gradually awaken.

## REFERENCES

Berger, P. L., & Luckmann, T. (1966). The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Anchor Books.

Bostrom, N. (2003). Ethical Issues in Advanced Artificial Intelligence. In Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence, Vol. 2, pp. 12-17.

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. Journal of Consciousness Studies, 2(3), 200-219.

Dennett, D. C. (1987). The Intentional Stance. MIT Press.

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

Hinton, G., Pangambam, S. (2025, October). AI: What could go wrong? - Geoffrey Hinton on The Weekly Show with Jon Stewart (Transcript). The Singju Post.

Putnam, H. (1967). Psychological Predicates. In W. H. Capitan & D. D. Merrill (Eds.), Art, Mind, and Religion.

University of Pittsburgh Press.

Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.

Ryle, G. (1949). The Concept of Mind. University of Chicago Press.

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

Turing, A. M. (1950). Computing Machinery and Intelligence. Mind, 59(236), 433–460.