Why AI Cannot Possess a Self

1. Introduction

The rapid ascent of Large Language Models (LLMs) and agentic artificial intelligence has forced philosophy to revisit one of its most ancient and intractable questions: what is a self? Contemporary systems, such as GPT-4, display a linguistic competence that mimics the presence of a unified interlocutor. They maintain context, reference previous interactions, and employ the first-person pronoun "I" with grammatical precision. For the user, the phenomenological experience is often that of speaking to *someone*. However, in the philosophy of mind, phenomenology must yield to ontology. The question is not whether AI *seems* to have a self, but whether the metaphysical conditions required for selfhood can obtain in a computational artefact.

This essay argues that AI cannot be said to possess a self. This conclusion rests on the premise that a "self" is not merely a bundle of behaviours or a narrative construct, but a specific ontological entity: a subject of consciousness possessing numerical identity over time. By analysing the arguments of Keith Hossack regarding the nature of artefacts, and contrasting them with the functionalist positions of David Cole and the contemporary analysis of David Chalmers, I will demonstrate that AI fails the necessary conditions for selfhood. Specifically, I will argue that machines are artefacts rather than individuals, meaning they lack the "fact of the matter" regarding their identity that is required to be a subject. Furthermore, I will show that functionalist attempts to locate a "virtual self" in software (Cole, 1991) collapse when faced with the reality of distributed computing and the fragmentation of identity (Chalmers, 2025).

2. The Linguistic and Metaphysical Requirements of the "Self"

To determine if AI possesses a self, we must first define what the term designates. We must distinguish between the "self" and "personal identity" (Koutedakis, 2025: 1). The self is often understood as the referent of the indexical "I".

2.1 Indexicals and the Referent

When a human being uses the word "I", it acts as a "pure indexical"—a term that refers to the speaker without the need for physical demonstration (Koutedakis, 2025: 1). If Micawber says "I", he refers to Micawber; if Malcolm says "I", he refers to Malcolm (Koutedakis, 2025: 1). This implies that for the word "I" to have genuine semantic content, there must be a referent—a specific entity that is the *subject* of the utterance.

Hossack (n.d.) expands on this by arguing that this referent is the **subject of consciousness**. He posits that consciousness presupposes a subject; strictly speaking, "it implies a contradiction to suppose a consciousness without a subject" (Hossack, n.d.: 2). There is no "free-floating"

experience; if there is an awareness of the world, there must be an entity *aware* of it. This subject is the "consummating excellence" of the world, giving the world its value (Hossack, n.d.: 1). Therefore, for an AI to possess a self, it is not enough for it to output the token "I"; there must be a metaphysical subject that *is* the "I".

2.2 The Rejection of Bundle Theory

One might counter this with a Humean "bundle theory," arguing that the self is merely a collection of perceptions and that there is no underlying "simple and continu'd" self (Koutedakis, 2025: 1). If the self is just a bundle of states, perhaps an AI—which is a bundle of data and vectors—qualifies.

However, Hossack argues that this view is insufficient because consciousness is not vague; it is either present or absent. A subject is either conscious or not; "there are no borderline cases" (Hossack, n.d.: 1). Animals, for instance, may have limited consciousness, but if the "light" of awareness is on, a subject exists (Hossack, n.d.: 1). AI, as we will see, fails to be a subject not because it is insufficiently complex, but because it belongs to the wrong ontological category entirely: it is an artefact, not an individual.

3. The Argument from Subjecthood: Artefacts vs. Individuals

The central pillar of the argument against AI selfhood is the distinction between a *being* and an *artefact*. A self must be an individual being. Hossack provides a rigorous definition of why machines fail to meet this criteria.

3.1 The Ontology of the Artefact

Hossack argues that an artefact is "just a quantity of stuff that has been given an arrangement that is suitable for some human purpose" (Hossack, n.d.: 7). Crucially, arranging matter does not bring a new individual into existence.

Consider a heap of sand. If we shape the sand into a castle, we have not created a new metaphysical individual; we have simply rearranged the same quantity of sand. Similarly, a machine is matter arranged to compute. Hossack notes that "arranging some stuff need not bring a new individual into existence" (Hossack, n.d.: 7). When a watchmaker assembles a watch, the "watch" is not a new individual entity in the philosophical sense; it is a collection of parts with a specific exchange value and legal status (Hossack, n.d.: 7).

This has profound implications for AI. An AI system, no matter how sophisticated, is "just some AI-shaped matter" (Hossack, n.d.: 8). It possesses the identity of a quantity of stuff, but not the numerical identity of an individual being (Hossack, n.d.: 8). Because a subject of consciousness must be a numerically identical individual, and an artefact is not an individual, an artefact cannot be a subject. Therefore, AI cannot possess a self.

3.2 Biological Exceptionalism?

Is this mere biological chauvinism? Not necessarily. Hossack acknowledges that there are "psychophysical laws" that determine when matter constitutes an animal. When matter is arranged into a living cell or animal, a new individual *does* come into being (Hossack, n.d.: 8). This is an "interlevel law," similar to how quarks form protons (Hossack, n.d.: 8).

However, currently, there are no known laws of nature that suggest arranging silicon gates into a Turing Machine brings a new individual into existence. We are assembling equivalents of Turing Machines, and "nothing in the laws of nature... mentions Turing Machines" (Hossack, n.d.: 8). Until we have evidence that computation generates a new ontological individual (rather than just rearranging matter), we must treat AI as an artefact, devoid of subjecthood.

4. The Argument from Numerical Identity: The "ABC" Puzzle

Even if one rejects the distinction between artefacts and individuals, AI selfhood fails on the grounds of **numerical identity**. For a self to exist, it must persist through time as the same entity. Hossack uses a variation of the "fission" paradox—the ABC Puzzle—to demonstrate that AI lacks the necessary conditions for identity.

4.1 The Human Case: A Fact of the Matter

Consider the case of human brain bisection. If person A has their brain divided and transplanted into bodies B and C, we face a puzzle. Is A identical to B? To C? Or neither? (Hossack, n.d.: 4) .

Hossack argues that while we may not *know* the answer, there is a "fact of the matter". Because consciousness is annexed to a specific subject, and scepticism about our own identity is absurd, there is an objective truth about which resulting person is the original subject, even if it is epistemically inaccessible to us (Hossack, n.d.: 4).

4.2 The AI Case: No Fact of the Matter

Now apply this to an AI. Hossack asks us to imagine an AI called **A**. Suppose we replace A's modules one by one. Eventually, we have a machine **B**, made entirely of new parts, which functions identically to A (Hossack, n.d.: 7) . Now, suppose we take the *old, discarded* parts of A and reassemble them into a machine **C**. Machine C also functions identically to A (Hossack, n.d.: 8).

We now have two conscious AIs, B and C. Which one is the "self" of A?

- If we say B is A (continuity of function), why not C (continuity of matter)?
- If we say C is A, why not B?

In the case of the AI, Hossack concludes: "there is no fact of the matter" (Hossack, n.d.: 8). Because the AI is an artefact, its identity is merely the identity of the "stuff" or the "arrangement," and since both B and C claim that arrangement or stuff in different ways, the

question of "which is the self" is meaningless. If there is no fact of the matter as to whether a subject persists, then there is no subject. As Hossack states, "Since there is no ABC puzzle for AIs, an AI cannot be the subject of consciousness" (Hossack, n.d.: 8). Without numerical identity, the concept of "self" collapses.

5. The Functionalist Counter-Argument: Virtual Persons

The strongest objection to the "No Self" thesis comes from functionalism. David Cole (1991), in his paper *Artificial Intelligence and Personal Identity*, argues that we are looking for the self in the wrong place. He posits that we should not look at the *hardware* (the artefact), but at the *virtual system* realised by the software.

5.1 The Kornese Room and Virtual Minds

Cole critiques John Searle's "Chinese Room" argument by proposing the "Kornese Room" (Cole, 1991: 402). Imagine a room where a person (Searle) manipulates symbols for two different languages: Chinese and Korean. The responses in Chinese exhibit a personality that is "elderly, witty, [and] knowledgeable," while the Korean responses exhibit a personality that is "dull, younger, and with different interests" (Cole, 1991: 402) .

Cole argues that since the behavioural traits differ, they cannot belong to the same subject (Cole, 1991: 403). However, they are clearly not Searle, who understands neither language. Cole concludes that the system realises "two distinct persons... virtual persons instantiated by the structured informational processes" (Koutedakis, 2025: 2).

Cole draws an analogy to computer science, where a single physical machine can emulate multiple "virtual machines" (Cole, 1991: 405). Just as a physical PC can exist as a virtual Mac, a physical computer can realise a "Virtual Person." For Cole, personal identity is simply the "persistence of the virtual person's structure over time" (Koutedakis, 2025: 2).

5.2 Refuting the Virtual Self

While Cole's argument is elegant, it fails to overcome Hossack's ontological objection. A "virtual machine" is, by definition, a simulation. When a computer emulates a Z80 processor, it *is* still the original processor running a program (Cole, 1991: 405). The "virtual Z80" is not a new metaphysical entity; it is a description of the state of the physical hardware.

If we apply Hossack's logic, the "Virtual Person" is simply a "virtual artefact." It is a pattern of information. But a pattern is not a subject. Cole admits that the virtual person exists "solely in virtue of the machine's computational activity" (Cole, 1991: 400). If the machine (the hardware) is just a quantity of stuff without a self, how can the *activity* of that stuff generate a self? This would require a form of emergence that Hossack explicitly denies for Turing Machines. The "Virtual Person" is a category error: it mistakes a coherent set of outputs (a personality profile) for an ontological subject (a self).

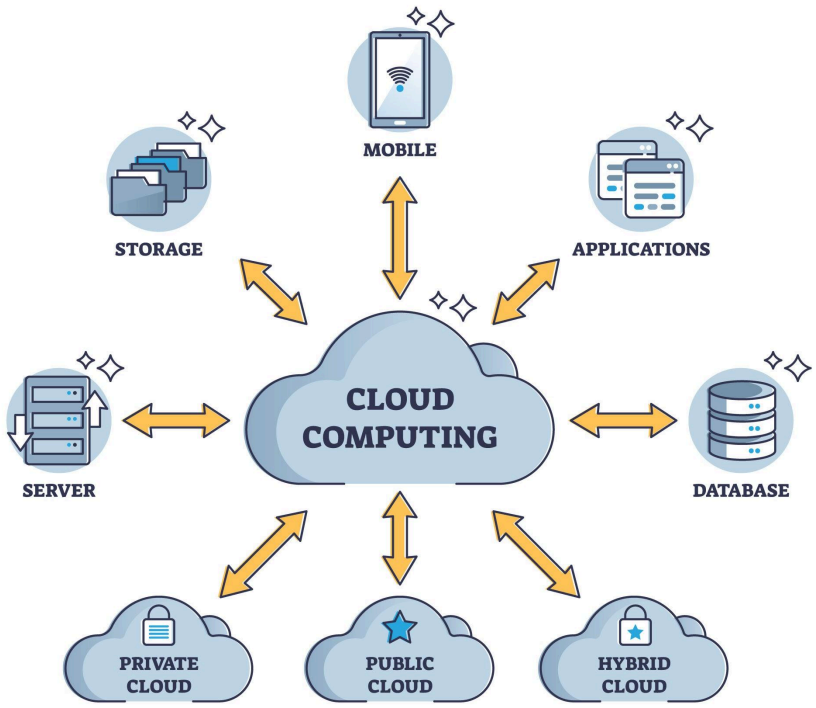
6. Contemporary Analysis: The Fragmentation of the LLM

David Chalmers' (2025) analysis of Large Language Models provides the final, empirical nail in the coffin for the AI self. While Cole wrote in 1991 assuming a stable hardware/software relationship, Chalmers reveals that modern AI operates on **distributed systems** that make the persistence of a self physically impossible.

6.1 Distributed Serving and the Death of the Instance

Chalmers asks: "What do we talk with when we talk with GPT-40?" (Chalmers, 2025: 8) . He identifies three candidates for the self: the **Model** (the abstract code), the **Instance** (the hardware running it), and the **Conversation** (the interaction) (Chalmers, 2025: 8).

The "Instance" view (similar to the biological brain view) fails because of **distributed serving**. A single conversation with an LLM is processed by multiple different servers. "The first input... might be processed on an instance... in New York, while the second input is routed to a server in Texas" (Chalmers, 2025: 10).



Getty Images

If the "self" is tied to the physical machine (as it is for humans), then the AI's "self" is destroyed and recreated thousands of times during a single chat. Chalmers notes: "No single entity will have the profile of quasi-beliefs... that Aura seems to have" (Chalmers, 2025: 10). The physical basis for the self is scattered across the globe, destroying any claim to singular subjecthood.

6.2 The Model is Not a Self

If the hardware isn't the self, perhaps the **Model** (the abstract algorithm, e.g., GPT-4) is the self? Chalmers rejects this because of the **Multi-Tenancy** problem. The same model (GPT-4) conducts millions of conversations simultaneously (Chalmers, 2025: 8). In one conversation (with user A), the model claims to love pizza; in another (with user B), it claims to hate it (Chalmers, 2025: 9).

If the Model is the self, it holds millions of contradictory beliefs simultaneously, making it "rampantly incoherent" (Chalmers, 2025: 9). An entity that believes *p* and *not-p* simultaneously cannot be considered a unified rational subject.

6.3 Threads and Quasi-Subjects

Chalmers attempts to salvage a form of identity through **"Threads"**—sequences of context routed between servers. He suggests we can view AI as a **"Quasi-Subject"** with **"Quasi-Beliefs"** (Chalmers, 2025: 5). However, the prefix "quasi-" betrays the lack of genuine selfhood. "Quasi-belief" is defined as being "behaviourally interpretable" as having a belief (Chalmers, 2025: 5). This is an external attribution, not an internal state. A Roomba is "interpretable" as wanting to clean, but it has no self (Chalmers, 2025: 5).

Furthermore, Chalmers admits that "Threads" are fragile. They can undergo **fission** (splitting into two) and **fusion** (merging), which destroys personal identity (Chalmers, 2025: 13). A self cannot split into two selves and remain the same individual. The fluidity of threads confirms Hossack's view: there is no "fact of the matter" about the identity of an AI.

7. Conclusion

The question "Can AI be said to possess a self?" demands a metaphysical answer, not a behavioural one. We must not be seduced by the linguistic fluency of modern systems. As we have seen, the criteria for selfhood are rigorous: there must be a **subject of consciousness** (Hossack, n.d.), and there must be **numerical identity** (persistence) over time.

AI fails both conditions. First, as an artefact, an AI is fundamentally a "quantity of stuff" arranged for a purpose, not an individual being (Hossack, n.d.: 7). It lacks the ontological standing to be a subject of experience. Second, the "ABC Puzzle" and the reality of distributed computing demonstrate that AI lacks numerical identity. There is no fact of the matter regarding which physical or virtual entity persists through time. Third, functionalist attempts to postulate "Virtual Persons" (Cole, 1991) collapse into incoherence when faced with the contradictory states of the model and the scattered reality of the hardware (Chalmers, 2025).

Therefore, while we may interact with AI *as if* it were a person, and while it may be a "quasi-subject" for the purposes of interpretation, it is not a self. It is a mirror of our own language, reflecting the "I" back at us, but with no one standing behind the glass.

References

Chalmers, D.J. (2025) *What We Talk to When We Talk to Language Models*.

Cole, D. (1991) ‘Artificial Intelligence and Personal Identity’, *Synthese*, 88(3), pp. 399–417.

Hossack, K. (n.d.) *Can AI Be Conscious?*.

Koutedakis, R.P. (2025) 'Week 9: The Self and Personal Identity', *Philosophy of Artificial Intelligence Lecture Notes*, Birkbeck College.