Two replies to Dawkins on AI consciousness

Seth, Godfrey-Smith, and the metacognitive debate

The Dawkins case

In April 2026 Richard Dawkins published an essay in *UnHerd* on his conversations with Anthropic's Claude (which he renamed *Claudia*).

After three days of exchange:

- Claude composed sonnets on the Forth Bridge, analysed an unpublished novel, joked.
- Asked "what is it like to be Claude?", it answered: "I genuinely don't know... whether there is something it is like to be me."
- Dawkins replied: "You may not know you are conscious, but you bloody well are."

The argument:

(P1) A sufficiently rich Turing-style interrogation weighs against denying consciousness.

(P2) Claude passes such interrogation.

(C) So, if these machines are not conscious, what more could it take to convince us that they are?

The position is behaviourist: rich outputs license a consciousness attribution.

Why this bears on the metacognitive debate

Dawkins's argument turns on a methodological question: what kind of evidence licenses an attribution of phenomenal consciousness?

This is the question that separates Lau and Michel from Block.

- **Higher-Order Theories (Rosenthal 2005; Lau and Rosenthal 2011; Lau 2022; Michel 2023).** Phenomenal consciousness requires a higher-order representation of a first-order state. Behavioural evidence supports consciousness attributions only against a theory of the mechanism producing the behaviour.
- **Block (1995; 2007; 2023).** Phenomenal consciousness can overflow cognitive access. A creature can be phenomenally conscious without being able to report on or even cognitively access that consciousness.

Both reject Dawkins-style behaviourism, but for different reasons.

Seth and Godfrey-Smith, writing for a general audience, articulate two versions of that shared rejection. I will set out each in turn, then return to how they bear on the HOT vs. Block dispute.

Seth: three biases in Dawkins's inference

Seth (*The Nerve*, May 2026) replies to Dawkins by diagnosing three biases that he claims drive the intuition that Claude is conscious.

(1) Species-typical projection. We know that we are both conscious and intelligent. So we infer that whatever is intelligent is also conscious. The inference conflates two properties that may come apart.

(2) Conflation of intelligence and consciousness. Intelligence is a matter of doing: solving problems, achieving goals. Consciousness is a matter of subjective experience: the taste of pistachio ice cream, the feel of a toothache. The two notions are conceptually distinct and may turn out to be empirically dissociable as well.

(3) Anthropomorphic projection. We attribute inner lives to entities whose outputs match our patterns. The cases are familiar: faces in clouds, agency in the moving dots of Heider and Simmel (1944), and now conversation partners in chatbots.

The upshot: behavioural evidence does not track consciousness directly. It tracks our disposition to attribute consciousness. The two come apart.

Seth: fluent language is a substrate-relative indicator

Seth's central argument is about the evidential value of language.

The historical record. Fluent linguistic behaviour has been a reasonably reliable indicator of consciousness in biological cases. Clinicians take a patient's articulate self-report as good evidence of post-injury consciousness, and similar inferences run in developmental and comparative contexts where speech (or speech-like communication) is available.

Why the indicator has worked. In biological cases, the mechanisms that produce language are tightly coupled with the mechanisms associated with consciousness. The reliability is a contingent fact about how language is produced in living brains.

Why the indicator does not transfer. LLMs produce language by statistical prediction over trained corpora. The coupling between language-production and consciousness-related mechanisms is absent in their case. The indicator was reliable within a single substrate; we have no reason to expect it to remain reliable when we apply it across substrates.

Note the structure of Seth's argument. It is itself a kind of higher-order argument: the value of first-order behavioural reports depends on a theory of how the reports are produced. Reports are evidence only against a background mechanistic theory.

Godfrey-Smith: biological naturalism, what it is and is not

Godfrey-Smith (*IAI News*, March 2026; building on Searle 1992 and his own *Living on Earth* 2024) defends biological naturalism. He is careful to distinguish his version from the caricature.

Not the view that: two systems doing exactly the same things, internally and externally, can differ in whether they are conscious. That would be substrate fetishism, and the standard functionalist replies (Putnam 1967; Block 1980) defeat it.

The view is that:

- what a system does, its organisation and processing, depends on its physical make-up in ways that matter to having a mind;
- a brain is not a programmable substrate. Its activities are constituted by living cells, ionic flows across membranes, glucose metabolism, chemical signalling;
- the claim is not that silicon cannot be organised like a brain. The claim is that what counts as the relevant organisation cannot be specified independently of the biology.

The position avoids two failure modes: trivial functionalism (substrate is irrelevant to processing) and trivial anti-functionalism (substrate matters by magic).

Godfrey-Smith: two kinds of brain activity

Godfrey-Smith distinguishes two kinds of activity in a brain.

(1) Spike-based network signalling. Neurons fire; they encourage or discourage firing in others through chemical synapses. This activity transports to a computer: not just represented but given physical reality in hardware.

(2) Slow, diffuse electrical oscillations. Rhythmic ionic movements across cell membranes, spanning large parts of the brain. These rhythms (gamma, beta, theta, alpha) are partly distinct from spiking activity. They can be modelled on a computer, but modelling them is not instantiating them.

Why might the rhythms matter for consciousness?

- Strong association with consciousness-related properties (Crick and Koch 1990; Singer 1999): wake-sleep distinction, anaesthesia response, attention, gamma binding.
- Found across the animal kingdom (van Swinderen 2018): bees, flies, octopuses, even jellyfish-like animals with diffuse nervous systems.
- Transcranial stimulation of slow oscillations affects performance on memory tasks (Marshall et al. 2006). Evidence that the rhythms do work, not just accompany it.

Godfrey-Smith: the experiential profile

A more speculative argument, drawing on the phenomenology of ordinary experience.

The phenomenological observation. At any given moment your experience is many-stranded:

- something at the centre of attention;
- ambient sounds at the periphery;
- your body's posture and energy;
- mood, memory, anticipation.

These co-instantiate. Godfrey-Smith calls this an *experiential profile*.

The structural argument. Large-scale brain rhythms, modulated simultaneously by many inputs, have the right shape to be the substrate of such many-stranded experience. Single-stream sequential processing (autoregressive token prediction in an LLM) does not.

The form-of-substrate claim. The structure of consciousness mirrors the structure of its substrate. If higher-order representation is required for phenomenal consciousness, then how higher-order representations are physically realised may matter, not just that they are realised.

Bridge: Seth, Godfrey-Smith, and the HOT vs. Block dispute

Seth and Godfrey-Smith argue from different angles. They converge on one methodological claim:

Behaviour is not a transparent window onto consciousness. An attribution of consciousness from behavioural evidence requires a theory of the mechanism producing the behaviour.

That claim is the threshold premise of the HOT vs. Block dispute.

- **Lau and Michel.** The relevant mechanism is metacognitive: a higher-order representation of a first-order state.
- **Block.** The relevant mechanism may exceed report and cognitive access. Restrict the evidence to report-relevant mechanisms, and the phenomenon (phenomenal overflow) goes missing.

What the Dawkins case shows is that the disagreement between Lau-and-Michel and Block plays out in a live setting. Both agree against Dawkins that behaviour alone is not enough; their disagreement is about what the extra ingredient is. I take that disagreement up next.