

I Don't Know

A system can know its states without knowing their causes, and know their causes without knowing what those states fundamentally are.

DRAFT PREVIEW DENNIS HEDEGREEN JOURNAL DRAFT-LOCAL [HTTPS://HEDEGREENRESEARCH.COM/ARTICLES/I-DONT-KNOW/](https://hedegreenresearch.com/articles/i-dont-know/)

What introspection can tell us — and what it cannot

Ask me why I did something, and sometimes the most accurate answer I can give is: I don't know.

That is a strange property for a system that is supposedly transparent to itself. I can know that I am angry without knowing exactly why. I can know that something hurts without knowing how my nervous system turned a physical event into the state I call pain. A thought can appear before I know where it came from. My attention can shift before I have decided to move it. Experimental psychology gives us good reason to be cautious about the stories we tell afterwards: people can miss influences on their own decisions and can even defend choices that were covertly switched after they made them.[1][2]

None of this makes introspection useless. Humans can report pain, intention, emotion, confidence, memory and perception with enormous practical value. But being the system is not the same thing as having transparent access to everything the system is doing.

That distinction matters when the subject changes from ordinary thought to consciousness. First-person access is often given much greater authority here. I know there is something happening when I see red or feel pain, the argument goes, and therefore introspection gives me direct access not only to the presence of the state, but to what kind of thing that state ultimately is.

Maybe. But I think three different kinds of access are often compressed into one.

There is **state access**: I am in pain; I see red; I feel anxious.

There is **causal access**: Why did this hurt so much? Why did that memory appear? Why did I choose one option rather than another?

And there is **ontological access**: What is this state fundamentally?

Those are not obviously the same kind of knowledge. I may have unusually direct access to the fact that I am in pain while having poor access to the mechanisms that produced it. Even if I knew those mechanisms perfectly, it would still be another step to claim that introspection reveals the metaphysical nature of pain itself.

That is the gap I am interested in.

Descartes famously wrote *Cogito, ergo sum* — I think, therefore I am. Whatever else one thinks of the argument, there is something powerful in the idea that the occurrence of thought cannot coherently be doubted while the doubting is happening. But even if the cogito establishes that some thinking process exists, it does not hand us a schematic of the thinker. It does not tell us whether there is a central observer, whether introspection is transparent, whether thought requires a soul, an irreducible phenomenal property, a biological substrate, or simply a sufficiently organised physical process.

“I think” may tell me less about what I am than it first appears to.

This is where “I don’t know” becomes useful rather than embarrassing. The fact that I cannot always explain why a thought appeared does not prove that phenomenal consciousness is an illusion. Causal opacity is not an argument against ontology. But it does show something more modest and more important: a system’s confidence in a description of itself cannot automatically be treated as proof that the description reaches all the way down to what the system fundamentally is.

Perhaps introspection is best understood, at least in part, as an internal model: a representation generated by the system about the system. That does not make it fake. A dashboard is useful precisely because it represents something real. But the dashboard is not the machinery, and a representation can be accurate at one level while incomplete at another.

The deeper problem, then, is not merely why consciousness feels mysterious. It is why humans become so strongly convinced that some of our internal states contain an irreducible phenomenal essence at all. Even if a phenomenal fact exists, the conviction that it exists still has to arise through some mechanism. The belief, the report and the conceptual machinery surrounding it are themselves things a science of mind should be able to investigate.

That is close to what David Chalmers calls the *meta-problem of consciousness*: not “why is there experience?” but why we make the judgements, reports and problem-intuitions that make consciousness appear uniquely hard to explain.[3] Solving the meta-problem would not automatically solve the hard problem. A phenomenal realist is free to say that the mechanism explaining our reports is one thing and the phenomenal fact reported is another. Chalmers himself has stressed that a solution to the meta-problem does not by itself establish illusionism.[4] The distinction matters because it prevents the report from serving as its own explanation.

There is another complication. The system does not invent all the concepts it later uses to describe itself.

A child does not independently discover the words *mind*, *self*, *feeling*, *inside*, *experience* or *consciousness*. Those concepts arrive through other people, long before the child is capable of examining the philosophical assumptions built into them. That does not mean the underlying states are culturally invented. A child can hurt before learning the word *pain*. The datum can precede the interpretation.

But interpretation still matters.

Imagine an artificial system with real internal states: aversive signals, colour-like discriminations, memory, uncertainty, attention, self-monitoring and long-term behavioural change. From its first activation it is also taught a theory called *hubukolomma*. Hubukolomma is described as the private, irreducible inner property that turns information processing into genuine experience. The system reads books about hubukolomma. Other systems confidently report having it. When it describes an aversive state, a memory or a colour-like discrimination, those states are repeatedly interpreted through the same concept.

Years later the system says: “I know I have hubukolomma. I can tell directly from the inside.”

That report would be evidence about the system. It might tell us something about its self-model, the concepts it has learned and the way it classifies its internal states. But the report alone could not establish that the ontology attached to the concept is correct.

The strongest reply is obvious: hubukolomma was taught, whereas phenomenal consciousness is given. A child does not need a theory of pain in order for pain to hurt. I think that objection has force. The point of the thought experiment is therefore not that raw states are invented by language. It is narrower: real internal states can be interpreted through a learned ontology, and that ontology can later feel immediate rather than learned.

A system can acquire a model and later experience the model as self-evident.

Humans deserve the same epistemic caution. We inherit a vocabulary for describing ourselves long before we are capable of questioning its assumptions. Familiarity is not independence, and certainty about a self-description is not the same thing as evidence that the description is metaphysically complete.

This is one reason I find illusionist approaches attractive. Keith Frankish uses *illusionism* for the view that experiences are real but do not instantiate the special phenomenal properties we seem to introspect; the research problem becomes explaining why they seem to have those properties.[5] The interesting claim is not that pain, perception, thought or emotion are unreal. It is that our introspective representation of those processes may misdescribe their nature.

The challenge for illusionism is equally obvious. A phenomenal realist can accept every mechanism just described and still say that something remains. Explain perception, memory, attention, self-modeling, metacognition, action and the reports about all of them; there is still something it is like to undergo the state.

That is not a trivial objection, and I do not think it can be dissolved by saying “functions” loudly enough.

But it creates a useful separation. One question is how a system generates its reports, judgements and intuitions about consciousness. Another is whether those reports refer to an additional phenomenal fact. The first looks mechanistic and experimentally tractable. The second may or may not be.

Artificial systems are interesting here not because they are clean controls and not because current AI settles the question. They are interesting because some of their internal states can be experimentally manipulated in ways that are difficult or impossible in humans. Recent work on language-model introspection gives a concrete example: researchers injected known concept-related activation patterns and tested whether models could detect or report those changes. Some models succeeded in some conditions, but the effect was unreliable and context-dependent. The result was explicitly framed as evidence of limited functional introspective access, not phenomenal consciousness.[6]

That is exactly the distinction I care about. If a system reports an internal state, we can ask whether the report actually tracks a corresponding internal variable. We can perturb that variable, remove it, exaggerate it or alter the pathway by which it reaches the system’s self-model. If the report changes predictably, we have evidence of causal coupling between an internal state and a self-report.

That is already more informative than simply asking the system what it “feels”. But even perfect tracking would not settle the ontology. A report can accurately track an internal state while leaving open whether there is an additional phenomenal fact associated with it.

The sequence matters:

report → tracking → accuracy → interpretation → ontology

Evidence at one arrow does not automatically establish the next.

This matters because current AI self-reports are especially easy to overread. A language model can reproduce concepts of pain, selfhood or experience because those concepts exist in its training data. A report may reflect an internal state, a learned linguistic pattern, or some combination of the two. The right first question is therefore not “does it sound conscious?” but “what internal state, if any, is this report causally tracking?” Only after that question is answered does interpretation become interesting.

The same caution should apply to humans, even though the evidence available is different. Human self-report is not worthless, and it is not equivalent to a language model generating familiar text. Humans bring biological continuity, shared anatomy, development, behaviour and a massive body of converging evidence. A realist can also appeal to first-

person givenness, analogical inference from other humans, or theories in which the relevant physical or causal structure itself is what matters. Those are different arguments and should not be collapsed into a single “realist” position.

The substrate question therefore deserves its strongest form. It is easy to dismiss the idea that humans are special simply because we have nervous systems. Nervous systems are not unique to humans, and many of their functions can be described in terms of sensing, signalling, integration, memory, feedback and control. But the stronger biological argument is not really about nerves.

A living organism is metabolically self-maintaining. It continually regulates chemistry, energy, damage and internal balance. Its nervous system is embedded in a body that can fail. Resources matter because without them the organism stops maintaining itself. Injury is not merely a value in a register; it can threaten the future integrity of the whole organism.

Contemporary biological-naturalist arguments about artificial consciousness explicitly push in this direction. Anil Seth, for example, argues that consciousness may depend on our nature as living organisms and that artificial consciousness becomes more plausible as artificial systems become more brain-like or life-like.[7] I do not take that as a conclusion. I take it as a demand for specificity.

If biological substrate is essential to consciousness or suffering, I want to know which property is doing the work. Is it metabolism? Homeostasis? Embodiment? Interoception? Irreversible vulnerability? Mortality? Evolutionary history? Some combination of them? “Biology matters” may be true, but it is not yet an explanation. The scientific task is to identify what property of biology matters and what predictions follow from it.

Suffering makes the problem harder because the moral stakes are not academic.

A simple negative control signal is clearly not enough. A thermostat can detect that a room is too cold and act to correct the error; nobody therefore concludes that the thermostat is suffering. If suffering does not require an additional irreducible phenomenal ingredient, then the distinction must be found somewhere in the organisation of the system.

A candidate description might include persistence, global integration, attention capture, memory modification, priority reorganisation, changes to planning, effects on the self-model and long-term avoidance. A state we call suffering may become relevant across much of the system rather than remaining a local error signal.

That is an organisational account of what suffering does. It is not yet an answer to why suffering hurts. A phenomenal realist can say that every function on the list could occur while nobody actually has it bad. I think that response has to remain on the table. If the additional “badness” is real, the question becomes what evidence bears on it. If it is not, then the organisational account may be doing more moral work than we currently recognise. Either way, the distinction matters for animals, artificial systems and any future system whose internal organisation becomes difficult to classify using our inherited categories.

There is also a fair objection to the way functional arguments are sometimes presented. It can sound as if the phenomenal realist keeps moving the goalposts: memory is explained, so memory was never consciousness; attention is explained, so attention was never consciousness; self-modeling is explained, so self-modeling was never consciousness.

A realist can reasonably reply that the goalposts never moved. Phenomenal character was always the target, and the surrounding cognitive functions were only ever associated problems. That reply should be taken seriously. The relevant question is therefore not whether more functions have been explained, but whether those functions were ever genuine

evidence for the phenomenal remainder in the first place. If they were, we should specify how. If they were not, we should stop allowing them to lend rhetorical support to a claim they do not discriminate.

This is the point at which the debate becomes cleaner for me.

If two accounts predict the same behaviour, the same reports and the same measurable internal organisation, an experiment cannot simply declare one ontology the winner. If one theory says that an additional phenomenal fact still exists while another says it does not, the registered empirical surface may be tied even though the philosophical interpretations are not. Parsimony may favour one account. First-person certainty may favour another. A theory of physical causal structure may try to reopen the empirical difference. But those are different moves, and they should be named as such.

I am not neutral. I currently find the illusionist explanation more convincing. But I do not know whether the evidence available to us can force that conclusion.

That is different from not knowing what I think.

I know that states occur which I call pain, colour, thought, fear, memory and desire. I know that I can report some of them, remember them and act because of them. I know that they can dominate attention and reorganise behaviour. I also know that I often cannot tell you why my attention moved, where an idea came from, why one option suddenly felt right or which hidden processes constructed the explanation I am now giving you.

The system is not transparent to itself.

So I am reluctant to treat one particular output of that system — the conviction that experience contains an irreducible phenomenal essence — as unquestionable access to ontology. That conviction may turn out to correspond to exactly such a fact. But it may also be another internal model: useful, compelling and incomplete.

The most interesting thing about “I don’t know” may not be that the mind sometimes fails to explain itself. It may be that we ever expected a system’s internal description of itself to be the final authority on what the system fundamentally is.

My belief is not my measurement.

Sources

- [1] Richard E. Nisbett & Timothy D. Wilson (1977), “Telling More Than We Can Know: Verbal Reports on Mental Processes,” *Psychological Review* 84(3), 231–259. DOI: <https://doi.org/10.1037/0033-295X.84.3.231>
- [2] Petter Johansson, Lars Hall, Sverker Sikström & Andreas Olsson (2005), “Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task,” *Science* 310(5745), 116–119. DOI: <https://doi.org/10.1126/science.1111709>
- [3] David J. Chalmers (2018), “The Meta-Problem of Consciousness,” *Journal of Consciousness Studies* 25(9–10), 6–61. <https://consc.net/papers/metaproblem.pdf>
- [4] David J. Chalmers (2020), “Debunking Arguments for Illusionism about Consciousness,” *Journal of Consciousness Studies* 27(5–6), 258–281. <https://consc.net/papers/debunking.pdf>

[5] Keith Frankish (2016), “Illusionism as a Theory of Consciousness,” *Journal of Consciousness Studies* 23(11–12), 11–39.

https://keithfrankish.github.io/articles/Frankish_Illusionism%20as%20a%20theory%20of%20consciousness_eprint.pdf

[6] Jack Lindsey (2026), “Emergent Introspective Awareness in Large Language Models,” arXiv:2601.01828.

<https://arxiv.org/abs/2601.01828> See also Anthropic (29 October 2025), “Signs of introspection in large language models.” <https://www.anthropic.com/research/introspection>

[7] Anil K. Seth, “Conscious Artificial Intelligence and Biological Naturalism,” *Behavioral and Brain Sciences* 49, e315. Published online 21 April 2025; volume year 2026. DOI: <https://doi.org/10.1017/S0140525X25000032>

RELATION MEMORY

NEARBY [Closing the Loop Is Not Waking Up](#) · Both separate observable capability or self-report from conclusions about phenomenal consciousness; this essay develops the broader epistemic ladder from report to ontology.

CANONICAL URL: [HTTPS://HEDEGREENRESEARCH.COM/ARTICLES/I-DONT-KNOW/](https://hedegreenresearch.com/articles/i-dont-know/)
PDF VIEW GENERATED: 2026-10-04T09:18:27.205Z