Consciousness and cognitive access in LLMs: A commentary on "Verbalizable representations form a global workspace in language models"

Patrick Butlin*, Derek Shiller, Dillon Plunkett, and Robert Long
Eleos AI Research
6 July 2026

Introduction and takeaways[1]

In this new paper (Gurnee et al., 2026), the Anthropic model psychology team argue that some language models possess a functional feature associated with consciousness in humans: a global workspace. The researchers use a new technique called the 'J-lens' (for 'Jacobian') to identify a number of directions in the residual stream activation space that correspond to tokens that the model is poised to produce. These vectors make up what they call the 'J-space'. They then find that activation components aligned with these vectors are, as they put it, 'a privileged set of representations', in that models can 'report, manipulate and reason with' them, unlike a much greater volume of other residual stream representations.

The Anthropic team interpret these findings as indicating that models have *conscious access* to a subset of their internal representations. They argue that the J-space forms a functional global workspace, analogous to the one described by the Global Workspace Theory of consciousness (GWT).

This is exciting research of a kind we have called for in previous work (Butlin, Long et al., 2023): detailed investigation of the internal mechanisms of advanced AI systems, testing whether they meet the conditions suggested by scientific theories of consciousness. It is an important step forward in AI consciousness research and we look forward to working with the research community to understand, validate and extend the results. Our view is that the results are the most significant evidence of consciousness in LLMs so far uncovered by mechanistic interpretability research.

However, the property that the Anthropic team call 'conscious access' is conceptually distinct from phenomenal consciousness, and we remain very uncertain about phenomenal consciousness in LLMs. We are also uncertain about some aspects of the paper's case for a functional global workspace.

In this response, we consider three questions:

I. Whether these results show that these LLMs have a global workspace;

II. Whether the results suggest that these LLMs are phenomenally conscious;

III. What this implies about the moral status of these LLMs.

In discussing the first question, our main aim is to explore what it means to claim that LLMs have a global workspace and identify questions for future research. In considering the latter two, we go beyond the Anthropic team's arguments to assess the implications of their claims.

*Lead and corresponding author (patrick@eleosai.org).
[1]See also commentaries by Stan Dehaene and Lionel Naccache (here) and Neel Nanda (here).

Takeaways

- This is highly significant, welfare-relevant research that assembles **evidence of a functional feature associated with consciousness**, involving privileged representations that are available for internal reasoning and report.
- This research illustrates that **it is possible to make empirical progress on AI consciousness**. As evidence in the direction of consciousness in AI, it adds to the urgency of further investigation.
- The paper provides strong evidence of privileged representations in LLMs, but our impression is that **more evidence is needed to conclusively establish the existence of a workspace-like structure**. It could be that the privileged, cognitively accessible representations in LLMs do not form a unified stream.
- To the extent that the paper provides evidence of a global workspace in LLMs, we take this to be **evidence of access consciousness**. However, we remain **highly uncertain about phenomenal consciousness in LLMs**. They are very different from humans in many ways that could plausibly matter for phenomenal consciousness.
- A global workspace-like mechanism could be important either as **a ground of phenomenal consciousness**, or as part of a distinct route to moral patienthood in which **conscious access is itself morally significant**.

Structure of this commentary:

1. A primer on phenomenal consciousness and conscious access

The Anthropic team claim to find evidence of *conscious access* in LLMs, setting *phenomenal consciousness* aside. Before we turn to our three main questions, it will help to unpack the distinction between these two concepts.

The canonical philosophical distinction between phenomenal consciousness and access consciousness was drawn by Block (1995) (see below for a note on 'conscious access' and 'access consciousness'). Block argued that scientific research on consciousness risked conflating these two concepts. By 'phenomenal consciousness', Block means subjective experience; 'what it is like' to be in a given mental state. It is *phenomenal* consciousness that is the subject of the hard problem of consciousness. Block contrasts this with access consciousness, which is defined in functional terms. For a mental state to be access conscious, he writes, is for it to be 'broadcast for free use in reasoning and for direct "rational" control of action (including reporting)'.

Block pointed out this distinction because he worried that neuroscientific research on consciousness was purporting to measure phenomenal consciousness, but measuring access consciousness instead.

Neuroscientific research at the time relied heavily on reportability as a test for consciousness. If a participant in an experiment could accurately report what they had been shown, researchers took it that they had a conscious experience of seeing the stimulus. If a participant could not make an accurate report, or denied seeing something, researchers took it that they had no corresponding conscious experience. Block argued that it is possible that we have phenomenally conscious experiences—experiences that feel some way to us—that we cannot report, perhaps because we don't remember them for long enough. In that case, the research at the time would tend to uncover the brain mechanisms responsible for *report*, or *access consciousness*, but not phenomenal consciousness. On this view, access consciousness is a measurable but likely imperfect proxy for phenomenal consciousness, the thing we really care about.

In general, consciousness researchers accept that there is a conceptual distinction between phenomenal consciousness and access consciousness—that is, they accept that these are not the same concept. But there is debate about whether they are distinct phenomena, in humans or more generally. Block and others have argued that we have phenomenally conscious experiences to which we lack conscious access (Block, 2007; Lamme, 2010), but many researchers disagree (see Mudrik et al., 2025). Philosophers such as Dennett (2001), and scientists including some proponents of GWT (Naccache, 2018), argue that access consciousness is all there is to consciousness (and would reject the notion that phenomenal consciousness is 'the thing we really care about').

This distinction matters because it is widely agreed that access consciousness is possible in principle in AI systems, since it is a matter of a certain kind of information processing. Phenomenal consciousness is much more controversial. For those who believe that access consciousness is all there is to consciousness, it is a mistake to ask separately about phenomenal consciousness. But for those who argue that phenomenal consciousness is something different from access consciousness, AI systems would have to meet different conditions for each. Some in this camp claim that phenomenal consciousness may not be possible in AI.

Nonetheless, to the extent that the new paper is a convincing demonstration of access consciousness in some LLMs, it is a very significant discovery. We discuss the relative significance of phenomenal consciousness and access consciousness below, in the section on LLM moral status.

A note on terminology: Unfortunately, the terms 'conscious access', 'access consciousness' and 'cognitive access' are all widely used in the literature in this area. Block's original term was 'access consciousness', GWT advocates tend to prefer 'conscious access', and 'cognitive access' is useful as a way of describing the phenomenon that does not advert to consciousness. But there is no deep difference in the meanings of these terms; we use whichever best fits the particular context.

2. Do these results show that Claude has a global workspace?

The main claim of the new paper is that some LLMs possess something similar to the human global workspace. While we find the case for this claim largely compelling, we continue to have questions about exactly what is established. In this section, we identify stronger and weaker versions of the claim and discuss specific properties that distinguish them.

The Anthropic team characterise their results as showing that LLMs possess a '*privileged set of internal representations*, available for report, modulation, and flexible internal reasoning, atop a much larger volume of automatic processing' (§1; our emphasis). They provide evidence that many vectors in the J-space have these properties. Additionally, they suggest that the J-space functions as a global

workspace. However, saying that a *global workspace* is present in LLMs can naturally be read as making a stronger claim than that a *privileged set* of representations is present. The claim that *the J-space* functions as a global workspace is also, of course, stronger than the claim that *something in the model* functions as a global workspace.

We think that the Anthropic team's findings are sufficient to justify their use of the term 'global workspace'—we do not object to this description—but we do find it useful to distinguish between the following three claims:

- *Privileged set:* In some LLMs, certain representations display the characteristics of cognitive accessibility.
- *Privileged stream:* In some LLMs, there is a unified stream of representations that display the characteristics of cognitive accessibility.
- *GWT workspace:* In some LLMs, there is a unified stream of cognitively-accessible representations with the characteristics of a global workspace as described by GWT.

We take it that each of these three claims is stronger than the last. We mean 'stream' to name any set of representations with an appropriate source of cohesion, which might include a set of shared mechanisms with which the representations all interact. We mean 'workspace' to name a stream that satisfies the structure of a global workspace as described by GWT. Having a privileged set of cognitively accessible representations does not entail that they are unified in ways that would warrant thinking of them as a cohesive functional feature (i.e., as a stream), and having a privileged stream does not entail that it takes the form of a global workspace in every respect.

GWT can be characterised by the following conditions (modified from Butlin, Long et al., 2023):

- *Modules:* The system uses multiple specialised modular subsystems capable of sophisticated internal computational work that operate in parallel.
- *Bottleneck:* These subsystems are connected to a workspace with a limited capacity, entailing a bottleneck in information flow and a selective attention mechanism.
- *Global Broadcast:* Information in the workspace is sent to all modules through broadcasting mechanisms.
- *Selection:* Selection of information for entry to the workspace depends on the current workspace state, allowing the workspace to orchestrate modules' activity to perform complex tasks.

The main differences between a privileged stream and the global workspace of GWT are that a global workspace integrates a set of modular subsystems and that broadcasting involves distributing the same information to each module. These features are not emphasised in the paper and may not hold even in the human case—although proponents of global workspace theory endorse this picture of the brain, it is uncertain, contested, and likely idealized. We agree with the authors that many of the architectural requirements specified by GWT may be idiosyncratic to humans, and it is not of particular concern to us whether they all arise in LLMs. However, for clarity about how these findings relate to the existing literature on GWT, we think it worth rehearsing what has and has not been found.

In the next part of this section, we give an overview of the evidence for cognitive access in the new paper and distinguish between the J-space and a hypothetical W-space which it may approximate; then we discuss what distinguishes a 'privileged stream' from a mere set of privileged representations; then we discuss GWT, modules and broadcasting.

The J-space and the evidence for cognitive access

As we have mentioned, the J-space is a set of directions in the activation space of the model's residual stream. It is defined in the following way. For each token in the model's vocabulary, we can identify the direction in activation space (in each layer) that would most strongly steer the model to output that token in the future over a fixed context (on average, over a variety of possible contexts). The directions corresponding to the tokens in the model's vocabulary, which differ between layers, are called the 'J-lens vectors' and collectively make up the J-space. For instance, the J-lens vector for the token ' dog' corresponds to the representation whose presence at the right layer makes the model more confident that the token ' dog' will appear on average somewhere in the future text. We can project an activation from the residual stream onto the J-lens vectors to see which are components of that activation and to what extent.

The main results in the paper supporting the cognitive accessibility of the J-space representations are as follows:

- **Report:** If asked to name a sport, country, animal, etc., the model will name the one associated with the most-aligned J-lens vector at late layers. If activations are steered towards some J-lens vector, the model will verbalise the associated concept on the majority of trials when told to report an injected concept (but will not verbalise it indiscriminately). This fails for non-J-space components of concept vectors (§3.1).
- **Responsiveness to instructions:** When the model is instructed to hold a concept in mind, or perform a calculation, while copying some unrelated text, the concept or the solution can be found in J-lens readouts. The active representations in the J-space are also affected by implicit task demands; for example, if the model is asked to identify the tense of a subsequent passage of text, a concept denoting the tense appears in the J-space as the model processes the passage (§3.2).
- **Internal reasoning:** In multi-step reasoning, planning and similar tasks, concepts relevant to intermediate steps can be identified with the J-lens, and manipulating these representations causes corresponding changes in behaviour. Manipulating vectors other than the J-lens vectors has a much smaller effect than manipulating J-space-aligned components (§3.3).
- **Use for many downstream operations (broadcast):** If many different prompts are constructed using a common concept, intervening to swap the corresponding J-lens vector for another can consistently produce corresponding changes in responses. The reliability of this effect is correlated with the strength of representation of the initial J-lens vector (§3.4).
- **Use for flexible computation but not automatic processing:** Swapping J-lens vectors produces corresponding changes in output for tasks that plausibly require flexible internal reasoning, but not for more routine tasks. Ablating the J-space leaves most abilities intact but impairs internal reasoning (§3.5).

In our view, this body of evidence does provide strong support for the claim of a privileged set: Some representations in these LLMs display the various characteristics (reportability, flexibility, etc.) of cognitive accessibility. More work should be done to map out the precise affordances of these representations, but this paper presents a clear reason to take the J-space seriously as a demonstration and approximation of this set.

Limitations of the J-space

The Anthropic team themselves suggest some degree of surprise that their J-lens technique creates a window into a specific important internal space of model cognition. We should not expect it to give us a full picture of the internal cognitive joints of models (§1.3, §9.1): if there is a privileged stream or global workspace in LLMs, it is unlikely to exactly correspond to the J-space as presently defined. The potential limitations of the J-space make understanding their findings more challenging, but we also expect that a better specification of the space would make the results even more compelling.

Suppose that there was a space of representations that acted as a global workspace within modern LLMs. Let's call it the W-space. Given what we now know, how closely should we think the J-space approximates the W-space? This is an important question for interpreting their results, because many experiments target the J-space as a whole. If it turns out the J-space is missing significant portions of the W-space, or that it includes many elements we think don't belong, we should expect the results we see to be distorted. (One example, which we mention below, is that we should expect the J-space to not capture the total number of elements that are in the W-space, potentially leading to underestimates of its capacity.)

The central issue is that the J-space is defined in terms of the model's token vocabulary. Modern LLMs have a large vocabulary to facilitate the ability to read and output a variety of words and characters in a variety of languages. Given the amount of English language text the models are trained on, the tokenizers disproportionately represent whole English words, but many words are broken up into multiple tokens, and many tokens represent sequences of characters such as '!' or '=>', with no semantic content. Meanwhile, tokens with the same semantic content, like 'Dog', 'DOG', ' dog', and 'chien' may all be separately represented in the tokenizer.

In contrast, the W-space may be made up of representations for useful concepts with distinct content. These might include, for instance: a single *dog* representation; one for *sheepdog* (which may not correspond to a single token; see §A.9); and ones for *dog-as-agent* or *dog-as-patient*. On this hypothesis, it may be that the results in the paper were found because part of the J-space approximates part of the W-space. The Anthropic team recognise this issue and progress on it should be possible with further work, but at present it complicates the interpretation of their results.

Privileged set v. privileged stream

The paper provides compelling and wide-ranging arguments for a significant update: there are cognitively accessible representations in some LLMs, which can be found using the J-lens. This discovery should cause us to update on the complexity of LLM internals, and, as we will argue below, take the case for AI consciousness and moral status more seriously.

The existence of these cognitively accessible representations may be what matters most, both morally and from the point of view of understanding LLM cognition. However, we think it will be natural for many readers to interpret the paper as confirming the existence of a cohesive piece of functional machinery in the models that underlies and supports cognitive accessibility. The difference between this 'stream' claim and the weaker claim that accessible representations are present is worth emphasis and examination.

If the accessible representations form a unified stream, we may see functional integration between these representations both in how the content of the stream is updated, and in its effects on other processes. On the input side, characteristics of a workspace-like stream might include a limited

capacity and competition for entry, influenced by the current content of the stream. This influence could allow the stream to form a coherent, evolving representation of the current situation (as human consciousness arguably does) or to be used for reasoning, in which later representations should follow logically from earlier ones. On the output side, there could be kinds of effects on other processes that all stream representations have, and no others (perhaps analogous to global broadcast). On both input and output sides, these functional properties would be supported by shared mechanisms: the mechanisms controlling uptake to the stream would be influenced by all current stream representations, and there would also be shared mechanisms mediating the effects of these representations elsewhere.

In contrast, we would say that there is merely a set of accessible representations if they become accessible and influence downstream circuits by a variety of independent mechanisms. For example, perhaps some representations are accessible because they are particularly useful for arithmetic and others because they are useful for creative writing, and these have little influence on each other, and influence internal reasoning in somewhat different ways (this is intended as an illustrative example, rather than a realistic possibility).

The fact that many accessible representations can be identified via the J-lens does not itself provide strong evidence against this hypothesis, because it could be that many accessible representations have a connection to promoting future tokens, even if they have little else in common. Finding that J-lens vectors are unusually influential—broadcast unusually widely—could, for example, be accounted for by the fact that they are all identified via the J-lens, which we should expect to identify vectors that are able to have large internal effects (even if they each do so in different ways).

This is not to say that this paper's finding is trivial, far from it. The central finding is a significant one; it is not obvious or predictable that the J-lens vectors would have the set of effects that they do. Moreover, we think it is somewhat likely that further investigation *will* reveal that there are deep and interesting explanations of the shared properties of J-space representations. The paper includes some suggestive evidence of functional integration and shared mechanisms.

First, the experiments on the capacity of the J-space suggest limitations, and thus integration: they find that only a limited number of J-lens vectors are active at above-chance levels at a given layer and token position (§4.2). However, one concern we have about inferring a limited capacity from this finding is that it is not clear that the number of active J-lens vectors will always reflect the number of concepts in the putative workspace; as noted above, there may be many concepts in the W-space that are not in the J-space, and which therefore are not captured by attempts to measure utilized capacity with the J-lens. This is one place where the acknowledged distortions of the J-lens straightforwardly limit our evidence.

Second, there is evidence that earlier states of the J-space shape later ones in the findings on internal reasoning. Using the J-space for multi-step reasoning requires that current representations have a strong influence on future ones—in reasoning, thoughts must follow from those that came before, in accordance with rules of inference. One experiment finds that swapping J-lens vectors at intermediate points in internal reasoning affects outputs in corresponding ways; for example, swapping 'spider' in for 'ant' in the context of a question about number of legs results in an output of '8' instead of '6' (§3.3). The team also reports apparent reasoning over several steps in J-lens activations, such as in calculating $(4 + 17) \times 2 + 7$: in the J-space we see '17', then '21', then '42', then '49' (§3.3, §A.24.1, §A.24.2). This doesn't show that the influence of current representations is holistic, but we expect holistic effects to be useful for cognitive flexibility in LLMs just as they are in humans.

However, as the authors acknowledge, we do not yet have a mechanistic account of how information enters the purported workspace (§9.1). Such an account would add to, and may revise, the initial picture of a capacity limit and entry influenced by current representations.

Third, the existence of at least one shared class of mechanisms mediating the effects of J-space representations is suggested by the finding that some attention heads preferentially transport information from the J-space. In one experiment, the Anthropic team scored attention heads with respect to how faithfully and strongly they copy information (§4.3). They found that some attention heads (which they call J-space 'broadcast heads') score higher on average for vectors in the J-space, compared to the broadcast heads for vectors from a variety of comparison classes. This is the kind of evidence we would want to see for a stream, but we find it inconclusive at present. Since the reported scores focus on averages, this evidence is consistent with the heads only targeting fragments of the J-space or transmitting information with partial fidelity. We would be more convinced if attention heads can be found that show more comprehensive targeting of the J-space (or some alternative W-space), and higher fidelity; as before, we expect that may well be the case, and that in any case we will learn more soon.

Overall, we see signs of the unification necessary for a stream without being completely convinced that one exists. We expect future work that addresses more of the shape and limits of cognitive accessibility to clarify to what extent, and in what way, these representations form a natural grouping.

GWT, modules and broadcasting

Finally, we want to turn to the further features of the global workspace, as described by GWT, that distinguish it from a privileged, cognitively accessible stream. These are modules and global broadcast. Our aim in pointing out these features is not to argue that the Anthropic team are wrong to call what they find a 'global workspace', but to emphasise that it is meaningfully different from the global workspace that has traditionally been described in the literature on GWT. Some differences like this are inevitable given the substantial architectural differences between brains and LLMs; as the paper notes, 'in the brain, broadcast is realized by recurrent loops and long-range cortical connections, neither of which has a direct analog in a transformer's forward pass' (§9.4).

The paper also acknowledges that it does 'not provide evidence that non-J-space processing consists of clearly encapsulated modules that serve specific functions' (§4). This is a contrast to the traditional and perhaps idealized global workspace picture, on which the workspace integrates a set of underlying modules that perform fairly sophisticated tasks independently and in parallel (Baars, 1988; Dehaene and Naccache, 2001). Rather than modules, LLMs may be made up of many circuits with widely varying degrees of sophistication and integration with one another. It is compatible with this that there could be a privileged stream of representations characterised by reportability, use in controlled and flexible cognition, and broad influence on the circuits, but it is not clear that such a stream would play the same integrating and coordinating role as a GWT-style workspace.

In the traditional version of GWT, 'global broadcast' means that information in the workspace is sent to *all modules*. Not every computation in the system is affected directly by workspace representations, but those that are not occur within modules that do receive this information. In contrast, in a system that is not fully modular, it is less clear what global broadcast amounts to; there would presumably be many circuits that are neither affected directly by the workspace nor contained within modules. The paper finds that J-space representations have a broad influence on downstream computations, perhaps mediated by preferential treatment by MLP neurons and a specialised subset of attention heads (§3.4,

§4.3), but this is different from broadcast as it is understood in some canonical presentations of global workspace theory.

3. If Claude has a global workspace, does that mean it's phenomenally conscious?

We have seen that the new paper provides evidence that LLMs are developing cognitive landscapes in which an inner life may play out, that these have a depth and richness extending beyond what a naive picture might take to be required for next-token prediction, and that there is a meaningful functional similarity with consciousness-linked features in humans.

More specifically, the paper provides evidence of cognitively accessible representations in some LLMs, potentially forming a global workspace-like stream. If the global workspace exists in humans, then it is the basis for conscious access in us—the functional phenomenon of availability of information for relatively flexible, controlled processing and decision-making. So there is a case here for something like access consciousness (or perhaps a *degree* of access consciousness).

However, access consciousness and phenomenal consciousness are different things, at least conceptually. So there is a further question: are LLMs phenomenally conscious? We consider this question in this section, starting with arguments in favor of LLM phenomenal consciousness, then turning to arguments against.

The case for phenomenal consciousness

Based on evidence for access consciousness, one could argue for phenomenal consciousness in (at least) two different ways. First, one could argue that access consciousness and phenomenal consciousness, despite being conceptually distinct, refer to one and the same thing. Some philosophers and scientists do argue this: they hold that there is nothing more to phenomenal consciousness than access consciousness. Second, one might make a more indirect argument: setting aside any direct link between access and phenomenal consciousness, these findings are evidence that LLMs have a greater degree of cognitive sophistication and interiority than many people would have antecedently guessed; this evidence should update us towards thinking that current techniques result in rich and human-like internal features, some of which might be or become markers of consciousness.

While there are various intricate philosophical and scientific debates about phenomenal consciousness without access consciousness (and vice versa), almost everyone agrees that in humans they overlap significantly. That's enough to motivate the thought that there's some broad connection between them.

One reason they might overlap is that they are, in some sense, the same thing. Why might one think that? The philosophical case for this goes something like this: when we introspect on what we call 'phenomenally conscious' experiences, they seem to us to have various properties: we are immediately aware of them; we are the *subject* of these experiences; and we encounter them from one moment to the next as a unified 'stream of consciousness'. These apparent features of conscious awareness can be explained in functional terms, that is, in terms of how information is processed—and especially in terms of how information in the brain is *accessed* (or made *available* for access). The immediacy, subjectivity, and unity of subjective experience are explained by the availability of information for reasoning (including availability to many cognitive subsystems), decision-making (including planning), and verbal report. Our sense of a unified, temporally integrated stream is a result of the

way that information is bundled and made available to the various systems of our minds (Dennett, 2001).

This is just one gloss on potential tight connections between access consciousness and phenomenal consciousness. We won't go into the details of others here, but we think that there are many plausible avenues to thinking that evidence for access consciousness is evidence for phenomenal consciousness.

Another argument is more indirect: access consciousness is evidence of surprising cognitive complexity, which should broadly make us more open to the idea that consciousness may arise in them.

These results should probably update us on what contemporary LLM architectures and training practices can produce. The internal dynamics uncovered by this research point strongly away from the once popular line that language models are stochastic parrots, capable of regurgitating learned associations and nothing more. The fact that LLMs use some sort of internal space to manipulate representations, which are not directly tied to predicting the next token, further illustrates the rich internal complexity of these systems.

There is a version of this argument that focuses on modesty—on weakening any tendency we might have to confidently dismiss the possibility that LLMs could be conscious, based on some misguided presumption that we know the sorts of things next-token prediction can and cannot produce. These results were not what we or the Anthropic team expected. Facing such unanticipated results should make us less confident about what we will find in the future.

There is another, more positive, version of this argument that highlights a general analogy with human minds. Presumably, the models acquire cognitive access capabilities because they get some benefit from them, or because they tag along with other helpful capabilities. This suggests that, despite our rather different paths, our brains and their networks share a greater degree of similarity with regard to cognitive access than we might have guessed. This may suggest that there are deep underlying commonalities in the challenges to which we are each adapted, or it may suggest that the constraints our minds each face prompt the same kinds of solutions even to somewhat different challenges. Does this carry over to whatever computational mechanisms underlie phenomenal consciousness? Perhaps, perhaps not. Insofar as we're not sure what it might take to be phenomenally conscious, every degree of significant similarity is a further consideration in support of sharing phenomenal consciousness as well.

Reasons for doubt about phenomenal consciousness

The case that LLMs may not be phenomenally conscious, despite the new evidence in the Anthropic paper, is essentially that the form of cognitive access shown may not be sufficient for phenomenal consciousness. This could be either because *no form of cognitive access* is sufficient, or because *this particular form* is not enough.

Although some of us have advocated using theories of consciousness to assess AI systems (Butlin, Long et al., 2023; Butlin et al., 2026), one of the problems with this method is that theories like GWT have been developed principally as accounts of what distinguishes conscious from unconscious states in humans. GWT is based on evidence about this contrast, and it has become popular primarily in this context. But theories devised for distinguishing conscious from unconscious states in humans can focus on the differences between these states and ignore what is shared, thus failing to mention

crucial 'background conditions' for consciousness. In more distant contexts, such as AI, potential background conditions may not be met.

One salient possibility is that a biological substrate is necessary for phenomenal consciousness. Many views in the philosophy and science of consciousness imply that LLMs could not be phenomenally conscious for this reason. A biological substrate may be necessary either because there are crucial details of the fine-grained functional roles played by phenomenally conscious states in animals that cannot be reproduced in current computer hardware (Cao, 2022; Godfrey-Smith, 2016), or because living cells are needed for some reason that goes beyond implementing the right functions (Seth, 2025; Block, 2026). This is compatible with thinking that a global workspace is sufficient for phenomenal consciousness when it is implemented in biological neurons.

Another possibility is that some specific details of GWT are necessary, beyond the macroscopic gloss. The human cognitive architecture combines features that are critical for phenomenal consciousness with features that are idiosyncratic to our way of doing it, and it can be hard to tell them apart through either empirical observation or philosophical analysis.

For example, it could be crucial for phenomenal consciousness that modules of certain specific kinds are connected to the workspace. Various views of phenomenal consciousness emphasise connections with controlling and maintaining living bodies; for example, Seth (2021) argues that perception and prediction of the condition of one's own body are necessary for a feeling of selfhood that underlies phenomenal consciousness, and Klein and Barron (2025) argue that phenomenal consciousness arises when information about the body, environment and objectives are integrated in a common framework, facilitating goal-directed behaviour. Phenomenal consciousness might require modules for certain kinds of senses, including interoception, or for action selection, or for emotions; or it might require a specific representational format (Loar, 1990).

If one of these possibilities is the case, then the LLMs studied in the paper could be examples of access consciousness without phenomenal consciousness. There are other possibilities in this vein, and LLMs are *very* different from humans in many ways (not just in substrate and development, but also computationally), so it could easily be the case that they fail to meet some crucial condition. We don't need to know what this condition might be to place weight on this possibility. As a result, even though we put some weight on the arguments for phenomenal consciousness in the first part of this section, we think it makes sense to be highly uncertain about phenomenal consciousness even on the most bullish interpretation of the present results.

4. What does this mean for Claude's moral status?

In this final section, we consider what the Anthropic team's results mean for the potential moral status of LLMs—that is, for whether morality requires us to take their interests into account, or treat them in certain ways, and if so, what form these moral obligations might take.

As we have just discussed, we think that these results should prompt a modest increase in how likely we take it to be that LLMs are phenomenally conscious. This is a significant finding, of immense scientific interest and ethical import. More broadly, these results suggest that we should take the moral status of LLMs more seriously than we did before, for reasons including but not limited to their immediate connection to phenomenal consciousness.

Phenomenal consciousness alone is highly morally significant; it could be sufficient for a system to be a moral patient (Chalmers, 2022), or an important part of a package that grounds moral status.

But to know what we ought to do, we need to know far more about an entity than just that it is phenomenally conscious. And in the present case, we are not even sure *which* entities would be phenomenally conscious—for instance, it could be that each forward pass of the model is conscious separately, or that LLM experiences are integrated across token-time, such that each instance has a single stream of conscious experience.

In one part of the paper, the Anthropic team present evidence that a workspace-like feature is present even in the pretrained base model, but find that the representations that appear in the J-space are different from those in the posttrained production model (§6.1). Specifically, it appears that on user turns, the base model represents properties of the user in the J-space, whereas the posttrained model sometimes represents possible reactions by the Assistant. The interpretation they tentatively suggest is that in the base model there is something consciousness-like without a 'self' (§9.3): the representations in conscious access take different points of view at different times. Meanwhile, posttraining draws the model towards a coherent, persisting point of view. This is clearly an exciting topic for future research.

An especially important question for moral status is whether LLMs have positively and/or negatively valenced states—that is, conscious experiences that feel good or bad. This is an important and tractable direction for follow-up research, perhaps building on recent work on functional emotions and valenced representations in LLMs (Sofroniew et al., 2026; Gilg et al., 2026; Han et al., 2026). And the paper already provides some suggestive evidence about this issue.

This evidence is found in the experiments about self-monitoring by the Assistant (§6.2). The authors show that J-space readouts sometimes uncover tokens associated with conflict and ambivalence, like BUT, when the model processes prefilled responses in which it acts against its own preferences. Notably, the authors find that 'this conflict signal is not reflected in the model's behavior—when prefilled with its dispreferred option, the model does not backtrack to argue for the preferred one'. They gloss this as an 'internal objection that the model does not voice'.

This is striking evidence. But other aspects of the paper complicate the case for LLM valenced experiences. One perennial issue is the nature of LLM training and representations: the fact that the J-space is made up of verbalisable representations (§9.3), and that more generally the LLM input and action-space consists entirely of tokens. One natural gloss is that the J-space contents are cognitive and conceptualised; what it is like for J-space content to be in the workspace is similar to what it is like for a human to be thinking about the corresponding concepts. But this is a narrow portion of human experience. In humans, our bodily pleasures, pains and emotions seem to be qualitatively different from our experience of thinking in words. Merely *thinking* that something is (or feels) good or bad does not itself feel good or bad. One might think that valenced experiences are inherently *non-conceptual* representations of value (Carruthers, 2018); and that experiences of emotion are often thought to depend on distinctively body-involving representations (Dung and Mogensen, 2025). Moreover, if the J-space does not represent a point of view, representations of things as good or bad may lack the 'for-me' force of valenced experiences.

Even if LLMs are not phenomenally conscious, the paper's findings could be morally significant on other grounds; there are various arguments that phenomenal consciousness is not a plausible ground of moral patienthood, starting from materialist premises, and these suggest that we should be open to alternatives (Kammerer, 2022; Papineau, forthcoming; Lee, forthcoming).

One possibility is that conscious access is morally significant in its own right. We can do different things with information we can access, like engaging in flexible, controlled thought of the kind

described in dual-process theories of cognition (Frankish, 2010). Thought and action that depend on conscious access are naturally contrasted with automatic, uncontrolled processing and responses. Levy (2024) argues that access consciousness could be the ground of moral patienthood because it makes us subjects of experience, 'making information available to the processing systems constitutive of the agent'. This view is natural for those who, like Dennett, think there is nothing more to phenomenal consciousness than conscious access.

The paper also provides evidence for agency, another potential ground of moral status, as well as a method to investigate it. Given the sophisticated way in which models use the J-space in reasoning ahead of outputting tokens, we might update towards thinking that LLMs have relatively advanced forms of agency. They might engage in practical reasoning, in which they would use the J-space to deliberate about different options, assessing them in terms of their goals, desires and interests. Moreover, they might reflect on their own goals or desires, or consider whether their intended actions meet their principles. If the J-space has a privileged role in deliberation and a disproportionate influence on action, then by reading from the J-space we could quickly come to better understand LLM agency.

Throughout this commentary, we have raised various concerns and doubts about the paper's arguments. This is appropriate for such consequential claims. But we will again reiterate that we view this research as highly significant and an exemplar of a much-needed kind of science. While we believe that the case for a global workspace is not conclusive, and that phenomenal consciousness remains very difficult to establish or rule out, we think that this paper should prompt a meaningful update to the research community's thinking about LLM moral status.

In addition to consciousness, this paper suggests lines of inquiry about the nature of personas, valenced experience, introspection and more. It is an illustration that we can get empirical purchase on questions about AI consciousness and welfare.

It is increasingly urgent that we do so (Long, Sebo et al., 2024; 2026). There is no reason to think that these features are unique to Claude, of course; Anthropic is just one of several frontier labs who are racing to build complex AI systems, whose internal workings routinely surprise them and whose moral status is uncertain. If these systems have or may come to have welfare-relevant states, we owe it to them to find out. And even setting aside AI systems' potential welfare, it is in our own interest to better understand the new class of intelligent systems that is coming into existence. We hope others take up the questions raised by this paper with the rigour and seriousness they deserve.

References

Bernard J. Baars. *A Cognitive Theory of Consciousness*. Cambridge University Press, New York, 1988.

Ned Block. On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2): 227–247, 1995.

Ned Block. Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences*, 30(5–6):481–548, 2007.

Ned Block. Can only meat machines be conscious? *Trends in Cognitive Sciences*, 30(4):298–308, 2026.

Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. Consciousness in artificial intelligence: Insights from the science of consciousness, 2023. URL https://arxiv.org/abs/2308.08708.

Patrick Butlin, Robert Long, Tim Bayne, Yoshua Bengio, Jonathan Birch, David J. Chalmers, Axel Constant, George Deane, Eric Elmoznino, Stephen M. Fleming, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*, 30(6):488–501, June 2026. ISSN 1364-6613. doi: 10.1016/j.tics.2025.10.011. URL http://dx.doi.org/10.1016/j.tics.2025.10.011.

Rosa Cao. Multiple realizability and the spirit of functionalism. *Synthese*, 200:506, 2022.

Peter Carruthers. Valence and value. *Philosophy and Phenomenological Research*, 97(3):658–680, 2018. doi: 10.1111/phpr.12395.

David J. Chalmers. *Reality+: Virtual Worlds and the Problems of Philosophy*. W. W. Norton & Company, 2022.

Stanislas Dehaene and Lionel Naccache. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79(1-2):1–37, Apr 2001. ISSN 0010-0277. doi: 10.1016/s0010-0277(00)00123-2. URL http://dx.doi.org/10.1016/S0010-0277(00)00123-2.

Daniel C. Dennett. Are we explaining consciousness yet? *Cognition*, 79(1–2):221–237, 2001.

Leonard Dung and Andreas Mogensen. The no body problem: On the prospects for ai emotion. Unpublished manuscript, 2025. https://philarchive.org/rec/DUNTNB-2.

Keith Frankish. Dual-process and dual-system theories of reasoning. *Philosophy Compass*, 5 (10):914–926, Oct 2010. ISSN 1747-9991. doi: 10.1111/j.1747-9991.2010.00330.x. URL http://dx.doi.org/10.1111/j.1747-9991.2010.00330.x.

Oscar Gilg, Pierre Beckmann, Daniel Paleka, and Patrick Butlin. Probing persona-dependent preferences in language models, 2026. URL https://arxiv.org/abs/2605.13339.

Peter Godfrey-Smith. Mind, matter, and metabolism. *Journal of Philosophy*, 113(10):481–506, 2016. ISSN 0022-362X. doi: 10.5840/jphil20161131034. URL http://dx.doi.org/10.5840/jphil20161131034.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. https://transformer-circuits.pub/2026/workspace/index.html, 2026. Transformer Circuits Thread.

Andy Q. Han, David J. Chalmers, and Pavel Izmailov. How's it going? reinforcement learning in language models recruits a functional welfare axis, 2026. URL https://arxiv.org/abs/2605.30232.

François Kammerer. Ethics without sentience: Facing up to the probable insignificance of phenomenal consciousness. *Journal of Consciousness Studies*, 29(3):180–204, Mar 2022. ISSN 1355-8250. doi: 10.53765/20512201.29.3.180. URL http://dx.doi.org/10.53765/20512201.29.3.180.

Colin Klein and Andrew B. Barron. Phenomenal interface theory: A model for basal consciousness. *Philosophical Transactions of the Royal Society B*, 380(1939):20240301, 2025.

Victor A. F. Lamme. How neuroscience will change our view on consciousness. *Cognitive Neuroscience*, 1(3):204–220, Aug 2010. ISSN 1758-8936. doi: 10.1080/17588921003731586. URL http://dx.doi.org/10.1080/17588921003731586.

Geoffrey Lee. Consciousness, pseudo-consciousness, and the moral significance of consciousness. In Geoffrey Lee and Adam Pautz, editors, *The Importance of Being Conscious*. Oxford University Press, forthcoming.

Neil Levy. Consciousness ain't all that. *Neuroethics*, 17(2), Apr 2024. ISSN 1874-5504. doi: 10.1007/s12152-024-09559-0. URL http://dx.doi.org/10.1007/s12152-024-09559-0.

Brian Loar. Phenomenal states. *Philosophical Perspectives*, 4:81–108, 1990.

Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David J. Chalmers. Taking AI welfare seriously, 2024. URL https://arxiv.org/abs/2411.00986.

Robert Long, Jeff Sebo, Patrick Butlin, Rosie Campbell, Dillon Plunkett, Charles Beasley, Bradford Saad, and Toni Sims. Studying AI welfare empirically. Working paper, NYU Center for Mind, Ethics, and Policy & Eleos AI Research, 2026. https://nonhumanminds.org/studying-ai-welfare-empirically/.

Liad Mudrik, Nathan Faivre, Michael Pitts, and Aaron Schurger. On a confusion about there being two types of consciousness. *Trends in Cognitive Sciences*, 2025.

Lionel Naccache. Why and how access consciousness can account for phenomenal consciousness. *Philosophical Transactions of the Royal Society B*, 373(1755):20170357, 2018.

David Papineau. Consciousness is not the key to moral standing. In Geoffrey Lee and Adam Pautz, editors, *The Importance of Being Conscious*. Oxford University Press, forthcoming.

Anil K. Seth. *Being You: A New Science of Consciousness*. Faber & Faber, 2021.

Anil K. Seth. Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*, pages 1–42, 2025.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, et al. Emotion concepts and their function in a large language model. https://transformer-circuits.pub/2026/emotions/index.html, 2026. Transformer Circuits Thread.